COMPARISON INEQUALITIES AND FASTEST-MIXING MARKOV CHAINS

JAMES ALLEN FILL AND JONAS KAHN

Abstract

We introduce a new partial order on the class of stochastically monotone Markov kernels having a given stationary distribution π on a given finite partially ordered state space $\mathcal{X}$. When $K \preceq L$ in this partial order we say that K and L satisfy a *comparison inequality*. We establish that if $K_1, \dots, K_t$ and $L_1, \dots, L_t$ are reversible and $K_s \preceq L_s$ for $s = 1, \dots, t$, then $K_1 \cdots K_t \preceq L_1 \cdots L_t$. In particular, in the time-homogeneous case we have $K^t \preceq L^t$ for every t if K and L are reversible and $K \preceq L$, and using this we show that (for suitable common initial distributions) the Markov chain Y with kernel K mixes faster than the chain Z with kernel L, in the strong sense that *at every time t* the discrepancy—measured by total variation distance or separation or L^2-distance—between the law of Y_t and π is smaller than that between the law of Z_t and π.

Using comparison inequalities together with specialized arguments to remove the stochastic monotonicity restriction, we answer a question of Persi Diaconis by showing that, among all symmetric birth-and-death kernels on the path $\mathcal{X} = \{0, \dots, n\}$, the one (we call it the *uniform chain*) that produces fastest convergence from initial state 0 to the uniform distribution has transition probability 1/2 in each direction along each edge of the path, with holding probability 1/2 at each endpoint.

We also use comparison inequalities

(i) to identify, when π is a given log-concave distribution on the path, the fastest-mixing stochastically monotone birth-and-death chain started at 0, and
(ii) to recover and extend a result of Peres and Winkler that extra updates do not delay mixing for monotone spin systems.

Among the fastest-mixing chains in (i), we show that the chain for uniform π is slowest in the sense of maximizing separation at every time.

1. Introduction and summary

A series of papers [6, 32, 5, 4] by Boyd, Diaconis, Xiao, and coauthors considers the following "fastest-mixing Markov chain" problem. A finite graph $G = (V, E)$ is given, together with a probability distribution π on V such that $\pi(i) > 0$ for every i; the goal is to find the fastest-mixing reversible Markov chain (FMMC) with stationary distribution π and transitions allowed only along the edges in E. This is a very important problem because of the use of Markov chains in Markov chain Monte Carlo (MCMC), where the goal is to sample (at least approximately) from π and the Markov chain is constructed only to facilitate generation of such

Date: September 27, 2011; revised March 25, 2012.

2010 *Mathematics Subject Classification.* 60J10.

Key words and phrases. Markov chains, comparison inequalities, fastest mixing, stochastic monotonicity, log-concave distributions, birth-and-death chains, ladder game.

Research of the first author supported by the Acheson J. Duncan Fund for the Advancement of Research in Statistics.

observations as efficiently as possible. As their criterion for FMMC, the authors minimize SLEM (second-largest eigenvalue in modulus—sometimes also called the absolute value of the "largest small eigenvalue"—defined as the absolute value of the eigenvalue of the one-step kernel with largest absolute value strictly less than 1), and they find the FMMC using semidefinite programming. (More precisely, [6, 5, 4] do this; [32] similarly deals with continuous-time chains and minimizes relaxation time. See these papers for further references; in particular, related work is found in [27].)

While most of the results in the series are numerical, both [5] and [4] contain analytical results. For the problem treated in [5] (which, as explained there, has an application to load balancing for a network of processors [10]), the graph G is a path (say, on $V = \{0, \dots, n\}$, with an edge joining each consecutively-numbered pair of vertices) with a self-loop at each vertex, π is the uniform distribution, and it is proved that the FMMC has transition probability $p(i, i+1) = p(i+1, i) = 1/2$ along each edge and $p(i, i) \equiv 0$ except that $p(0, 0) = 1/2 = p(n, n)$. [We will call this the *uniform* chain $U = (U_t)_{t=0,1,\dots}$.]

The mixing time of a Markov chain can indeed be bounded using the SLEM, which provides the asymptotic exponential rate of convergence to stationarity. (See, e.g., [1] for background and standard Markov chain terminology used in this paper.) But the SLEM provides only a surrogate for true measures of discrepancy from stationarity, such as the standard total variation (TV) distance, separation (sep), and L^2-distance. For the path problem, for example, Diaconis [personal communication] has wondered whether the uniform chain might in fact minimize such distances after any given number of steps (when, for definiteness, all chains considered must start at 0). In this paper we show that this is indeed the case: The uniform chain is truly fastest-mixing in a wide variety of senses. Consider any $t \geq 0$. What we show, precisely, is that, for any birth-and-death chain[1] X having symmetric transition kernel on the path and initial state 0, the probability mass function (pmf) π_t of X_t majorizes the pmf σ_t of U_t. (A definitive reference on the theory of majorization is [21].) We will show using this that four examples of discrepancy from uniformity that are larger for X_t than for U_t are (i) $L^p(\pi)$-distance for any $1 \leq p \leq \infty$ (including the standard TV and L^2 distances); (ii) separation; (iii) Hellinger distance; and (iv) Kullback–Leibler divergence.

The technique we use to prove that π_t majorizes σ_t is new and remarkably simple, yet quite general. In Section 2 we describe our method of comparison inequalities. We show (Corollary 2.5) that if two Markov semigroups satisfy a certain *comparison inequality* at time 1, then they satisfy the same comparison inequality at all times t. We also show, in Section 3 (see especially Corollary 3.3), how the comparison inequality can be used to compare mixing times—in a variety of senses—for the chains with the given semigroups.

In Section 4 we show that, in the context of the above *path-problem* (of finding the FMMC on a path), if one restricts either (i) to monotone chains, or (ii) to even times, then the uniform chain satisfies a favorable comparison inequality in comparison with any other chain in the class considered. Somewhat delicate arguments (needed except in the case of L^2-distance) specific to the path-problem allow us to remove the parity restriction from the conclusion that the uniform chain is fastest. (See Theorem 4.3.) Further, comparisons between chains—even

[1] Arbitrary holding is allowed at each state.

time-inhomogeneous ones—other than the fastest U can be carried out with our method by limiting attention either to monotone kernels or to two-step kernels. Indeed, our Proposition 2.4 rather generally provides a new tool for the notoriously difficult analysis of time-inhomogeneous chains, whose nascent quantitative theory has been advanced impressively in recent work of Saloff-Coste and Zúñiga [28, 29, 31, 30].

In Section 5 (see Theorem 5.1), we generalize our path-problem result as follows. Let π be a log-concave pmf on $\mathcal{X} = \{0, \dots, n\}$. Among all *monotone* birth-and-death kernels K, the fastest to mix (again, in a variety of senses) is K_π with (death, hold, birth) probabilities given by

$$q_i = \frac{\pi_{i-1}}{\pi_{i-1} + \pi_i}, \qquad r_i = \frac{\pi_i^2 - \pi_{i-1}\pi_{i+1}}{(\pi_{i-1} + \pi_i)(\pi_i + \pi_{i+1})}, \qquad p_i = \frac{\pi_{i+1}}{\pi_i + \pi_{i+1}}.$$

(This reduces to the uniform chain when π is uniform.)

In Section 6 we revisit the birth-and-death problems of Sections 4–5 in terms of an alternative notion of mixing time employed by Lovász and Winkler [20]. Consider, for example, the path-problem of Section 4. For every even value of n the uniform chain is fastest-mixing in their sense, too. But, perhaps somewhat surprisingly, for every odd value of n the uniform chain is *not* fastest-mixing in their sense; we identify the chain that is.

In Section 7 we discuss a simple “ladder” game, where the class of kernels is a certain subclass of the symmetric birth-and-death kernels considered in Section 4.

In Section 8 we show how comparison inequalities can recover and extend (among other ways, to certain card-shuffling chains) a Peres–Winkler result about slowing down mixing by skipping (“censoring”) updates of monotone spin systems. (This is an example of comparison inequalities applied to time-inhomogeneous chains.)

2. Comparison inequalities

In this section we introduce our new concept of *comparison inequalities*. Consider a pmf $\pi > 0$ on a given finite partially ordered state space $\mathcal{X}$. We utilize the usual $L^2(\pi)$ inner product

$$\langle f, g \rangle \equiv \langle f, g \rangle_\pi := \sum_{i \in \mathcal{X}} \pi(i) f(i) g(i); \tag{2.1}$$

if a matrix K is regarded in the usual fashion as an operator on $L^2(\pi)$ by regarding functions on $\mathcal{X}$ as column vectors, then the $L^2(\pi)$-adjoint of K (also known as the time-reversal of K, when K is a Markov kernel) is K^* with $K^*(i,j) \equiv \pi(j)K(j,i)/\pi(i)$. Reversibility with respect to π for a Markov kernel K is simply the condition that K is self-adjoint.

Let $\mathcal{K}$, $\mathcal{M}$, and $\mathcal{F}$ denote the respective classes of (i) Markov kernels on $\mathcal{X}$ with stationary distribution π, (ii) nonnegative non-increasing functions on $\mathcal{X}$, and (iii) kernels K from $\mathcal{K}$ that are stochastically monotone (meaning that $Kf \in \mathcal{M}$ for every $f \in \mathcal{M}$). Note for future reference that the identity kernel I always belongs to $\mathcal{F}$, regardless of π. Define a *comparison inequality* relation $\preceq$ on $\mathcal{K}$ by declaring that $K \preceq L$ if $\langle Kf, g \rangle \le \langle Lf, g \rangle$ for every $f, g \in \mathcal{M}$, and observe that $K \preceq L$ if and only if the time-reversals K^* and L^* satisfy $K^* \preceq L^*$.

Remark 2.1. (a) Clearly,

(i) to verify a comparison inequality $K \preceq L$ by establishing $\langle Kf, g\rangle \le \langle Lf, g\rangle$, it is sufficient to take f and g to be indicator functions of down-sets (i.e., sets D such that $y \in D$ and $x \le y$ implies $x \in D$) in the partial order; and
(ii) if a comparison inequality holds, then the condition that f and g be nonnegative can be dropped, if desired.

(b) There is an important existing notion of *stochastic ordering* for Markov kernels on $\mathcal{X}$: We say that $L \le_{\mathrm{st}} K$ if $Kf \le Lf$ entrywise for all $f \in \mathcal{M}$. It is clear that $L \le_{\mathrm{st}} K$ implies $K \preceq L$ when K and L belong to $\mathcal{F}$. But in all the examples in this paper where we prove a comparison inequality, we *do not* have stochastic ordering. This will typically be the case for interesting examples, since the requirement for distinct $K, L \in \mathcal{F}$ to have the same stationary distribution makes it difficult (though not impossible) to have $L \le_{\mathrm{st}} K$.

Remark 2.2. The relation $\preceq$ defines a partial order on $\mathcal{K}$. Indeed, reflexivity and transitivity are immediate, and antisymmetry follows because one can build a basis for functions on $\mathcal{X}$ from elements f of $\mathcal{M}$, namely, the indicators of principal down-sets (i.e., down-sets of the form $\langle x\rangle := \{y : y \le x\}$ with $x \in \mathcal{X}$). A proof from first principles is easy.[2]

We list next a few basic properties of the comparison relation $\preceq$ on $\mathcal{K}$, showing that the relation is preserved under passages to limits, mixtures, and direct sums. The proofs are all very easy. Note also that the class $\mathcal{F}$ of stochastically monotone kernels with stationary distribution π is closed under passages to limits and mixtures, and also under (finite) products, but not under general direct sums as in part (c).

Proposition 2.3.

(a) *If $K_t \preceq L_t$ for every t and $K_t \to K$ and $L_t \to L$, then $K \preceq L$.*

(b) *If $K_t \preceq L_t$ for $t = 0, 1$ and $0 \le \lambda \le 1$, then*

$$(1-\lambda)K_0 + \lambda K_1 \preceq (1-\lambda)L_0 + \lambda L_1.$$

(c) *Partition $\mathcal{X}$ arbitrarily into subsets $\mathcal{X}_0$ and $\mathcal{X}_1$, and let each $\mathcal{X}_i$ inherit its partial order and stationary distribution from $\mathcal{X}$. For $i = 0, 1$, suppose $K_i \preceq L_i$ on $\mathcal{X}_i$. Define the kernel K (respectively, L) as the direct sum of K_0 and K_1 (resp., L_0 and L_1). Then $K \preceq L$.*

The following proposition, showing that $\preceq$ is preserved under product for stochastically monotone reversible kernels, is the main result of this section.

Proposition 2.4 (**Comparison Inequalities**). *Let $K_1, \dots, K_t$ and $L_1, \dots, L_t$ be reversible [i.e., $L^2(\pi)$-self-adjoint] kernels all belonging to $\mathcal{F}$, and suppose that $K_s \preceq L_s$ for $s = 1, \dots, t$. Then the product kernels $K_1 \cdots K_t$ and $L_1 \cdots L_t$ (and their time-reversals) belong to $\mathcal{F}$, and $K_1 \cdots K_t \preceq L_1 \cdots L_t$.*

[2] We need only show that the indicator function $\mathbf{1}_{\{x\}}$ of any singleton $\{x\}$ can be written as a linear combination of indicator functions of principal down-sets. But this can be done recursively by starting with minimal elements x and then using the identity

$$\mathbf{1}_{\{x\}} = \mathbf{1}_{\langle x\rangle} - \sum_{y<x} \mathbf{1}_{\{y\}}, \quad x \in \mathcal{X}.$$

The application to time-homogeneous chains is the following immediate corollary.

Corollary 2.5. *If $K, L \in \mathcal{F}$ are reversible and $K \preceq L$, then for every t we have $K^t, L^t \in \mathcal{F}$ and $K^t \preceq L^t$.*

Remark 2.6. As we shall see from examples, the applicability of our new technique of comparison inequalities is limited (i) by the monotonicity requirement for membership in $\mathcal{F}$ and (ii) by the extent to which $\mathcal{F}$ is ordered by $\preceq$. But restriction (i) in the choice of kernel has the payoff (among others) that the perfect simulation algorithms (see [33] for background) Coupling From The Past [25, 24, 26, 34] and FMMR (Fill–Machida–Murdoch–Rosenthal) [15, 16] can often be run efficiently for monotone chains. Restriction (ii) needs to be explored thoroughly for interesting and important examples. This paper treats a few examples, in Sections 4 (especially 4.1), 5, and 8. For discussion about the relation between our comparison-inequalities technique and existing techniques for comparing mixing times of Markov chains, see Remark 3.5 below.

The remainder of this section is devoted to the proof of Proposition 2.4, which we will derive as a consequence of an extremely simple, but—as far as we know—new, matrix-theoretic result, Proposition 2.7.

The general setting is this. We are given a positive vector $\pi \in \mathbf{R}^n$ and define the $L^2(\pi)$ inner product as at (2.1). We are also given a set (not necessarily a subspace) $W \subseteq \mathbf{R}^n$. Let $M_n(\mathbf{R})$ denote the collection of n-by-n real matrices. Define

$$\mathcal{F} := \{\text{matrices } A \in M_n(\mathbf{R}) \text{ for which } W \text{ is invariant}\}.$$

(This of course means that a real matrix A belongs to $\mathcal{F}$ if and only if $Aw \in W$ for every $w \in W$.) Define a (clearly reflexive and transitive) relation $\preceq$ on $M_n(\mathbf{R})$ by declaring that $A \preceq B$ if

$$\langle Ax, y\rangle \le \langle Bx, y\rangle \quad \text{for every } x, y \in W.$$

We observe in passing (i) that $A \preceq B$ if and only if $A^* \preceq B^*$ and (ii) that the relation $\preceq$ may fail to be antisymmetric (but this will present no difficulty).

Proposition 2.7. *Let $A_1, A_2, B_1, B_2 \in M_n(\mathbf{R})$. Suppose that A_2 and B_1^* both belong to $\mathcal{F}$. If $A_1 \preceq B_1$ and $A_2 \preceq B_2$, then $A_1 A_2 \preceq B_1 B_2$.*

Proof. Given $x, y \in W$, we observe

$$\begin{aligned}\langle A_1 A_2 x, y\rangle &\le \langle B_1 A_2 x, y\rangle \quad \text{because } A_2 x, y \in W \text{ and } A_1 \preceq B_1\\ &= \langle A_2 x, B_1^* y\rangle\\ &\le \langle B_2 x, B_1^* y\rangle \quad \text{because } x, B_1^* y \in W \text{ and } A_2 \preceq B_2\\ &= \langle B_1 B_2 x, y\rangle,\end{aligned}$$

as desired. □

The third (Corollary 2.10) of the following four easy corollaries of Proposition 2.7 implies Proposition 2.4 immediately, by setting $W = \mathcal{M}$ and observing that the set of Markov kernels with stationary distribution $\pi > 0$ is closed under both multiplication and adjoint. (Similarly, Corollary 2.5 is a special case of Corollary 2.11.)

Corollary 2.8. *Let A_1, A_2, B_1, B_2 be matrices all belonging to $\mathcal{F}$ with adjoints all belonging to $\mathcal{F}$, and suppose that $A_1 \preceq B_1$ and $A_2 \preceq B_2$. Then the matrices $A_1 A_2$ and $B_1 B_2$ and their adjoints all belong to $\mathcal{F}$, and $A_1 A_2 \preceq B_1 B_2$.*

Proof. This is immediate from the definition of $\mathcal{F}$ and Proposition 2.7. □

Corollary 2.9. *Let $A_1, \dots, A_t$ and $B_1, \dots, B_t$ be matrices all belonging to $\mathcal{F}$ with adjoints all belonging to $\mathcal{F}$, and suppose that $A_s \preceq B_s$ for $s = 1, \dots, t$. Then the matrices $A_1 \cdots A_t$ and $B_1 \cdots B_t$ and their adjoints all belong to $\mathcal{F}$, and $A_1 \cdots A_t \preceq B_1 \cdots B_t$.*

Proof. This follows by induction from Corollary 2.8. □

Corollary 2.10. *Let $A_1, \dots, A_t$ and $B_1, \dots, B_t$ be self-adjoint matrices all belonging to $\mathcal{F}$, and suppose that $A_s \preceq B_s$ for $s = 1, \dots, t$. Then the matrices $A_1 \cdots A_t$ and $B_1 \cdots B_t$ (and their adjoints) belong to $\mathcal{F}$, and $A_1 \cdots A_t \preceq B_1 \cdots B_t$.*

Proof. This is immediate from Corollary 2.9. □

Corollary 2.11. *Let A and B be self-adjoint matrices both belonging to $\mathcal{F}$, and suppose that $A \preceq B$. Then, for every $t = 0, 1, 2, \dots,$ the matrices A^t and B^t (are self-adjoint and) belong to $\mathcal{F}$ and $A^t \preceq B^t$.*

Proof. This is immediate from Corollary 2.10 by taking $A_s \equiv A$ and $B_s \equiv B$. □

3. Consequences of the comparison inequality, some via majorization

In this section we focus on time-homogeneous chains and show how comparison inequalities can be used to compare mixing times—in a variety of senses—for chains with the given semigroups. As we shall see in Section 3.3, a useful tool in moving from a comparison inequality to a comparison of mixing times will be the use of basic results from the theory of majorization.

3.1. Comparison inequalities and domination.

Recall from Section 2 that $\mathcal{F}$ denotes the class of stochastically monotone Markov kernels on a given finite partially ordered state space $\mathcal{X}$ that have a given π as stationary distribution. Our next result (Proposition 3.2) gives conditions implying that if a comparison inequality holds between reversible kernels $K, L \in \mathcal{F}$, then the univariate distributions of the corresponding Markov chains satisfy corresponding stochastic inequalities. The proposition utilizes the following definition.

Definition 3.1. Let (Y_t) and (Z_t) be stochastic processes with the same finite partially ordered state space. If for every t we have $Y_t \geq Z_t$ stochastically, i.e.,

$$\mathbf{P}(Y_t \in D) \leq \mathbf{P}(Z_t \in D) \text{ for every down-set } D \text{ in the partial order,} \tag{3.1}$$

then we say that Y *dominates* Z.

Proposition 3.2. *Suppose that $K, L \in \mathcal{F}$ are reversible and satisfy $K \preceq L$. If Y and Z are chains (i) started in a common pmf $\hat{\pi}$ such that $\hat{\pi}/\pi$ is non-increasing and (ii) having respective kernels K and L, then Y dominates Z.*

Proof. By Corollary 2.5 for every t we have $K^t, L^t \in \mathcal{F}$ and $K^t \preceq L^t$. The desired result now follows easily. □

3.2. **TV, separation, and L^2-distance.** Domination (recall Definition 3.1) is quite useful for comparing mixing times in at least three standard senses.

If d is some measure of discrepancy from stationarity, then in the following theorem we write "Y mixes faster in d than does Z" for the strong assertion that at every time t we have d smaller for Y than for Z.

Corollary 3.3. *Consider (not necessarily reversible) Markov chains Y and Z with common finite partially ordered state space $\mathcal{X}$, common initial distribution $\hat{\pi}$, and common stationary distribution π. Assume that $\hat{\pi}/\pi$ is non-increasing.*

(a) **[total variation distance]** *Suppose that Y dominates Z and that the time-reversal of Y is stochastically monotone. Then Y mixes faster in TV than does Z.*

(b) **[separation]** *Adopt the same hypotheses as in part* (a). *Then Y mixes faster in separation than does Z; equivalently, any fastest strong stationary time for Y is stochastically smaller (i.e., faster) than any strong stationary time for Z.*

(c) **[L^2-distance]** *Assume that Y and Z are reversible. Suppose, moreover, that the two-step chain (Y_{2t}) dominates (Z_{2t}) and is stochastically monotone. Then Y mixes faster in L^2 than does Z.*

Proof. All three results are simple applications of the domination inequality (3.1) [which, in the case of part (c), is guaranteed only for even values of t] or its immediate extension to expectations of non-increasing functions. We make the preliminary observation that $\mathbf{P}(Y_t = i)/\pi(i)$ is non-increasing in i for each t; indeed, writing K for the kernel of Y we have

$$\frac{\mathbf{P}(Y_t = i)}{\pi(i)} = \sum_j \frac{\hat{\pi}(j) K^t(j,i)}{\pi(i)} = \sum_j K^{*t}(i,j) \frac{\hat{\pi}(j)}{\pi(j)}, \tag{3.2}$$

so the non-increasingness claimed here follows from the monotonicity assumptions about $\hat{\pi}/\pi$ and K^*.

(a) Choosing D in (3.1) to be the down-set $D = \{i : \mathbf{P}(Y_t = i)/\pi(i) > 1\}$ we find

$$\mathrm{TV}_Y(t) = \mathbf{P}(Y_t \in D) - \pi(D) \le \mathbf{P}(Z_t \in D) - \pi(D) \le \mathrm{TV}_Z(t).$$

(b) We first observe

$$\mathrm{sep}_Y(t) = \max_i \left[1 - \frac{\mathbf{P}(Y_t = i)}{\pi(i)} \right] = 1 - \frac{\mathbf{P}(Y_t = x_1)}{\pi(x_1)}$$

for some maximal element x_1 in $\mathcal{X}$. Therefore, choosing $D = \mathcal{X} \setminus \{x_1\}$ we find

$$\begin{aligned} \mathrm{sep}_Y(t) &= 1 - \frac{\mathbf{P}(Y_t = x_1)}{\pi(x_1)} \\ &\le 1 - \frac{\mathbf{P}(Z_t = x_1)}{\pi(x_1)} \le \max_i \left[1 - \frac{\mathbf{P}(Z_t = i)}{\pi(i)} \right] = \mathrm{sep}_Z(t). \end{aligned}$$

(c) Using routine calculations suppressed here, one finds that the squared $L^2(\pi)$-distance (of the density with respect to π) from stationarity for Y_t equals

$$\begin{aligned} \sum_i \pi(i) \left[\frac{\mathbf{P}(Y_t = i)}{\pi(i)} - 1 \right]^2 &= \sum_{j'} \left[\sum_j \hat{\pi}(j) K^{2t}(j,j') \right] \frac{\hat{\pi}(j')}{\pi(j')} \; - \; 1 \\ &= \sum_{j'} \mathbf{P}(Y_{2t} = j') \frac{\hat{\pi}(j')}{\pi(j')} \; - \; 1. \end{aligned}$$

But $\hat{\pi}/\pi$ is non-increasing and $Y_{2t} \geq Z_{2t}$ stochastically; so this last expression does not exceed

$$\sum_{j'} \mathbf{P}(Z_{2t} = j') \frac{\hat{\pi}(j')}{\pi(j')} \; - \; 1 = \sum_{i} \pi(i) \left[\frac{\mathbf{P}(Z_t = i)}{\pi(i)} - 1 \right]^2 ,$$

which is the desired conclusion. □

We remark in passing that a very similar proof as for Corollary 3.3(b) gives the analogous result for the measure of discrepancy

$$\max_i \left[\frac{\mathbf{P}(Y_t = i)}{\pi(i)} - 1 \right],$$

and so we also have the analogous result for the two-sided measure

$$\max_i \left| \frac{\mathbf{P}(Y_t = i)}{\pi(i)} - 1 \right|. \tag{3.3}$$

Remark 3.4. [L^2-distance revisited] We have limited the statement of Corollary 3.3(c) to reversible chains for simplicity. The same proof shows, more generally, for each t that if (i) K and L are (not necessarily reversible) kernels with common stationary distribution π, (ii) $\hat{\pi}/\pi$ is non-increasing, and (iii) $\hat{\pi} K^t K^{*t} \geq \hat{\pi} L^t L^{*t}$ stochastically, then the $L^2(\pi)$-distance from stationarity for Y_t does not exceed that for Z_t, where the chains Y and Z have respective kernels K and L and common initial distribution $\hat{\pi}$. Assuming (i)–(ii), for the stochastic inequality (iii) here it is sufficient that K and L and their time-reversals K^* and L^* are all stochastically monotone and $K \preceq L$.

Remark 3.5. [concerning eigenvalues] (a) if K and L are ergodic reversible kernels in $\mathcal{F}$ (with a common stationary distribution π) and we have the comparison inequality $K \preceq L$, then the SLEM for K is no larger than the SLEM for L. This follows rather easily from Proposition 3.2 and Corollary 3.3(c) using the spectral representations of the kernels and the ample freedom in choice of the common initial distribution $\hat{\pi}$ such that $\hat{\pi}/\pi$ is non-increasing. We omit further details.

(b) There are several existing standard techniques for comparing mixing times of Markov chains, such as the celebrated eigenvalues-comparison technique of Diaconis and Saloff-Coste [9], but none give conclusions as strong as those available from combining Proposition 3.2 and Corollary 3.3. On the other hand, comparison of eigenvalues requires verifying far fewer assumptions than needed to establish $K, L \in \mathcal{F}$ and a comparison inequality $K \preceq L$, so our new technique is much less generally applicable.

3.3. **Other distances via majorization.** We now utilize ideas from majorization; see [21] for background on majorization and the concept of Schur-convexity used below. For the reader's convenience we recall that, given two vectors v and w in $\mathbf{R}^N$ (for some N), we say that v *majorizes* w if (i) for each $k = 1, \dots, N$ the sum of the k largest entries of w is at least the corresponding sum for v and (ii) equality holds when $k = N$. A function ϕ with domain $D \subseteq \mathbf{R}^N$ is said to be *Schur-convex on* D if $\phi(v) \geq \phi(w)$ whenever $v, w \in D$ and v majorizes w. Thus, given any two pmfs ρ_1 and ρ_2 on $\mathcal{X}$, if ρ_1 majorizes ρ_2, then for any Schur-convex function ϕ on the unit simplex (i.e., the space of pmfs) we have $\phi(\rho_1) \geq \phi(\rho_2)$. Examples of Schur-convex functions are given in Example 3.8 below; for each of those examples, the inequality $\phi(\rho_1) \geq \phi(\rho_2)$ can be interpreted as "ρ_2 is closer to π than is ρ_1".

The next proposition describes one important case where we have majorization and hence can extend the conclusions "Y mixes faster in d than does Z" of Corollary 3.3 to other measures of discrepancy d. Note the additional hypothesis, relative to Corollary 3.3, that π is non-increasing.

Proposition 3.6. *Consider (not necessarily reversible) Markov chains Y and Z with common finite partially ordered state space $\mathcal{X}$, common initial distribution $\hat{\pi}$, and common stationary distribution π. Suppose that both π and $\hat{\pi}/\pi$ are non-increasing. Suppose, moreover, that Y dominates Z and that the time-reversal of Y is stochastically monotone. Then, for all t, the pmf π_t of Z_t majorizes the pmf σ_t of Y_t.*

Proof. As noted just above (3.2), the ratio $\mathbf{P}(Y_t = i)/\pi(i)$ is non-increasing in i; since $\pi(i)$ is also non-increasing, so is the product $\mathbf{P}(Y_t = i)$. Hence for each $k \leq |\mathcal{X}|$ there is a down-set D_k such that $\mathbf{P}(Y_t \in D_k)$ equals the sum of the k largest values of $\mathbf{P}(Y_t = i)$. Since Y dominates Z, inequality (3.1) implies that, for all t, the pmf π_t of Z_t majorizes the pmf σ_t of Y_t. (This can be equivalently restated in language introduced in [13]: Z_t is coarser than Y_t, for all t.) □

Corollary 3.7. *Suppose that $K, L \in \mathcal{F}$ are reversible and satisfy $K \preceq L$, and that their common stationary distribution π is non-increasing. If Y and Z are chains (i) started in a common pmf $\hat{\pi}$ such that $\hat{\pi}/\pi$ is non-increasing and (ii) having respective kernels K and L, then, for all t, the pmf π_t of Z_t majorizes the pmf σ_t of Y_t.*

Proof. The desired conclusion follows immediately upon combining Propositions 3.2 and 3.6. □

Example 3.8. In this example we show when π is uniform in Proposition 3.6 (or Corollary 3.7), then Y mixes faster than does Z in more senses than TV, separation, and L^2.

Write N for the size of the state space $\mathcal{X}$. Then each of the following six functions is Schur-convex on the unit simplex in $\mathbf{R}^N$:

$$\begin{aligned}
\phi_1(v) &:= \Big[N^{p-1} \sum_i \left|v_i - N^{-1}\right|^p\Big]^{1/p} \quad \text{(for any } 1 \leq p < \infty\text{)},\\
\phi_2(v) &:= \max_i |N v_i - 1|,\\
\phi_3(v) &:= \max_i (1 - N v_i),\\
\phi_4(v) &:= \tfrac{1}{2} \sum_i \left(v_i^{1/2} - N^{-1/2}\right)^2,\\
\phi_5(v) &:= N^{-1} \sum_i \ln\left(\frac{1/N}{v_i}\right),\\
\phi_6(v) &:= \sum_i v_i \ln(N v_i);
\end{aligned}$$

in [21, Chapter 3], see Sections I.1, I.1, A.2, I.1.b, D.5, and D.1, respectively. Therefore, if ρ_1 majorizes ρ_2, then ρ_2 is closer to π than is ρ_1 in each of the following six senses (where here π is uniform and we have written the discrepancy from π for a generic pmf ρ):

(i) L^p**-distance**

$$\left[\sum_i \pi(i)\left|\frac{\rho(i)}{\pi(i)}-1\right|^p\right]^{1/p},$$

for any $1 \le p < \infty$;

(ii) L^∞**-distance**

$$\max_i \left|\frac{\rho(i)}{\pi(i)}-1\right|,$$

also called **relative pointwise distance**;

(iii) **separation**

$$\max_i \left[1-\frac{\rho(i)}{\pi(i)}\right];$$

(iv) **Hellinger distance**

$$\frac{1}{2}\sum_i \pi(i)\left[\sqrt{\frac{\rho(i)}{\pi(i)}}-1\right]^2;$$

(v) the **Kullback–Leibler divergence**

$$D_{KL}(\pi\|\rho) = -\sum_i \pi(i)\ln\left[\frac{\rho(i)}{\pi(i)}\right];$$

(vi) the **Kullback–Leibler divergence**

$$D_{KL}(\rho\|\pi) = \sum_i \rho(i)\ln\left[\frac{\rho(i)}{\pi(i)}\right].$$

Of course, the L^2-distance considered in Corollary 3.3(c) is the special case $p = 2$ of example (i) here, and the TV distance of Corollary 3.3(a) amounts to the special case $p = 1$. Relative pointwise distance was also treated earlier without use of majorization at (3.3).

4. Fastest mixing on a path

We now specialize to the path-problem. Let K be any symmetric birth-and-death transition kernel on the path $\{0, 1, \ldots, n\}$, and denote $K(i, i+1) = K(i+1, i)$ by p_i [except that $K(0, 0) = 1 - p_0$ and $K(n, n) = 1 - p_{n-1}$]; for example, when $n = 3$ we have

$$K = \begin{bmatrix} 1-p_0 & p_0 & 0 & 0 \\ p_0 & 1-p_0-p_1 & p_1 & 0 \\ 0 & p_1 & 1-p_1-p_2 & p_2 \\ 0 & 0 & p_2 & 1-p_2 \end{bmatrix}.$$

In this section we first show, in Sections 4.1–4.2, that if one restricts attention either

(i) to monotone chains, or
(ii) to even times,

then the uniform chain U with kernel K_0 where $p_i \equiv 1/2$ satisfies a favorable comparison inequality in comparison with the general K-chain, and we can apply all the results of Section 3. Then, in Section 4.3, we show that the parity restriction in (ii) can be removed to conclude that the uniform chain is, among all symmetric birth-and-death chains, closest to uniformity (in several senses) at all times. In this section and the next we make use of the general observation that a discrete-time birth-and-death chain with kernel K on $\mathcal{X} = \{0, 1, \dots, n\}$ is monotone if and only if

$$K(i, i+1) + K(i+1, i) \le 1 \text{ for } i = 0, \dots, n-1. \tag{4.1}$$

Before we separate into the two cases (i) and (ii) for the path-problem, let us note that if f is the indicator of the down-set $\{0, 1, \dots, \ell\}$, then Kf satisfies

$$(Kf)_j = \begin{cases} 1 & \text{if } 0 \le j \le \ell - 1 \\ 1 - p_\ell & \text{if } j = \ell \\ p_\ell & \text{if } j = \ell + 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

(with $p_n = 0$); hence if g is the indicator of the down-set $\{0, 1, \dots, m\}$, then

$$\langle Kf, g \rangle = \frac{1}{n+1} \times \begin{cases} m+1 & \text{if } 0 \le m \le \ell - 1 \\ \ell + 1 - p_\ell & \text{if } m = \ell \\ \ell + 1 & \text{if } \ell + 1 \le m \le n. \end{cases} \tag{4.3}$$

4.1. **Restriction to monotone chains.** Applying (4.1), our symmetric kernel K is monotone if and only if $p_i \le 1/2$ for $i = 0, \dots, n-1$. Among all such choices, it is clear that (4.3) is minimized when $K = K_0$. From Remark 2.1(i) it therefore follows that $K_0 \preceq K$ and hence from Section 3 (especially Corollary 3.7 and Example 3.8) that K_0 is fastest-mixing in several senses.

Remark 4.1. In fact, from (4.3) we see that monotone symmetric birth-and-death kernels K are monotonically decreasing in the partial order $\preceq$ with respect to each p_i.

4.2. **Restriction to even times.** In the present setting of symmetric birth-and-death kernel, note that our restriction (simply to ensure that K is a kernel) on the values $p_i > 0$ is that $p_i + p_{i+1} \le 1$ for $i = 0, \dots, n-1$. It is then routine to check that K^2 is (like K) reversible and (perhaps unlike K) monotone. Indeed, if f is the indicator of the down-set $\{0, 1, \dots, \ell\}$, then $K^2 f$ satisfies

$$(K^2 f)_j = \begin{cases} 1 & \text{if } 0 \le j \le \ell - 2 \\ 1 - p_{\ell-1} p_\ell & \text{if } j = \ell - 1 \\ 1 - 2p_\ell + 2p_\ell^2 + p_{\ell-1} p_\ell & \text{if } j = \ell \\ 2p_\ell - 2p_\ell^2 - p_\ell p_{\ell+1} & \text{if } j = \ell + 1 \\ p_\ell p_{\ell+1} & \text{if } j = \ell + 2 \\ 0 & \text{otherwise,} \end{cases} \tag{4.4}$$

which is easily checked to be non-increasing in j.

Suppose now that g is the indicator of the down-set $\{0, 1, \dots, m\}$. Then using (4.4) we can calculate, and subsequently minimize over the allowable choices of $p_0, \dots, p_{n-1}$, the quantity $\langle K^2 f, g \rangle$ by considering three cases:

(a) Suppose $m = \ell$. Then

$$(n+1)\langle K^2 f, g \rangle = \ell + (1 - p_\ell)^2 + p_\ell^2$$

is minimized (regardless of value ℓ) when $p_i = 1/2$ for $i = 0, \dots, n-1$.

(b) Suppose ℓ and m differ by exactly 1, say, $m = \ell + 1$. Then

$$(n+1)\langle K^2 f, g\rangle = \ell + (1 - p_\ell) + p_\ell(1 - p_{\ell+1}) = \ell + 1 - p_\ell p_{\ell+1}$$

is minimized (regardless of ℓ) when $p_i = 1/2$ for $i = 0, \dots, n-1$.

(c) Suppose ℓ and m differ by at least 2, say, $m \geq \ell + 2$. Then

$$(n+1)\langle K^2 f, g\rangle = \ell + (1 - p_\ell) + p_\ell + 0 = \ell + 1$$

doesn't depend on the choice of the vector $\mathbf{p}$.

From Remark 2.1(i) it therefore follows that $K_0^2 \preceq K^2$ and hence (from Section 3) that K_0^2 is fastest-mixing in several senses. Specifically:

$$\text{for all } \textit{even } t, \text{ the pmf } \pi_t \text{ of } X_t \text{ majorizes the pmf } \sigma_t \text{ of } U_t \tag{4.5}$$

if X and U have respective kernels K and K_0 and common non-increasing initial pmf $\hat{\pi}$. Further, when we consider all symmetric birth-and-death chains started in state 0, it follows from Corollary 3.3(c) that the chain with kernel K_0 is fastest-mixing in L^2 (without the need to restrict to even times, nor to monotone chains).

Remark 4.2. From the above calculations we see more generally that if K and $\widetilde{K}$ are two symmetric birth-and-death kernels and for every i we have

$$\left|p_i - \tfrac{1}{2}\right| \geq \left|\tilde{p}_i - \tfrac{1}{2}\right| \quad \text{and} \quad p_i p_{i+1} \leq \tilde{p}_i \tilde{p}_{i+1},$$

then $\widetilde{K}^2 \preceq K^2$.

4.3. **Removal of parity restriction.** Throughout this subsection all chains are assumed to start at state 0, even when we do not explicitly declare so. The main result of this section is the following theorem, which extends (4.5) to *all* times $t = 0, 1, 2, \dots$ and therefore demonstrates (by Example 3.8) that the uniform chain is fastest to mix in a variety of senses.

Theorem 4.3. *Let X be a birth-and-death chain with state space $\mathcal{X} = \{0, 1, \dots, n\}$ and symmetric kernel, and let U be the uniform chain. Suppose that both chains start at 0, and let π_t (respectively, σ_t) denote the probability mass function of X_t (respectively, U_t). Then*

$$\pi_t \text{ majorizes } \sigma_t \text{ for all } t.$$

Let X have kernel K as described at the outset of Section 4. Let Π_t and Σ_t denote the cumulative distribution functions (cdfs) corresponding to π_t and σ_t, respectively: for example,

$$\Sigma_t(j) := \sum_{i=0}^{j} \sigma_t(i) = \mathbf{P}(U_t \leq j).$$

From Section 4.2 we already know that if t is even then

$$\Pi_t(i) \geq \Sigma_t(i) \text{ for all } i, \tag{4.6}$$

because then π_t majorizes σ_t and both pmfs are non-increasing.

We build to the proof of Theorem 4.3 by means of a sequence of lemmas. We start with a few results about the uniform chain.

Lemma 4.4.

(a) *For every time t, the pmf σ_t is non-increasing on its domain $\{0, \dots, n\}$.*

(b) *The distribution "evolves by steps of two", depending on parity: for $i = 0, \dots, n-1$ we have*

$$\sigma_t(i) = \sigma_t(i+1) \qquad \text{if } t+i \text{ is odd.}$$

(c) *For every time t, the cdf Σ_t is concave (at integer arguments):*

$$2\Sigma_t(i) \ge \Sigma_t(i+1) + \Sigma_t(i-1), \qquad i \ge 0. \tag{4.7}$$

(d) *The inequality (4.7) is equality if $i \ge 0$ and t and i have opposite parity:*

$$2\Sigma_t(i) = \Sigma_t(i+1) + \Sigma_t(i-1) \qquad \text{if } t+i \text{ is odd.}$$

Proof. (a) This was proved in a more general setting just above (3.2).

(b) We use induction on t. The base case $t = 0$ is obvious $(0 = 0)$.

Using the induction hypothesis at the second equality, we conclude, when t and $i \in \{1, \dots, n-1\}$ have opposite parity, that

$$\sigma_t(i) = \tfrac{1}{2}\left[\sigma_{t-1}(i-1) + \sigma_t(i+1)\right] = \tfrac{1}{2}\left[\sigma_{t-1}(i) + \sigma_{t-1}(i+2)\right] = \sigma_t(i+1).$$

Similarly, when t is odd we have

$$\sigma_t(0) = \tfrac{1}{2}\left[\sigma_{t-1}(0) + \sigma_t(1)\right] = \tfrac{1}{2}\left[\sigma_{t-1}(0) + \sigma_{t-1}(2)\right] = \sigma_t(1).$$

(c) We first remark that it is well known that (4.7) is indeed equivalent to concavity of Σ_t at integer arguments. We then need only note that (4.7) is merely a rewriting of the monotonicity in part (a). Indeed,

$$\begin{aligned} 2\Sigma_t(i) &= \Sigma_t(i+1) + \Sigma_t(i-1) + \sigma_t(i) - \sigma_t(i+1) \\ &\ge \Sigma_t(i+1) + \Sigma_t(i-1). \end{aligned} \tag{4.8}$$

(d) Again using the equality at (4.8), this is merely a rewriting of the "steps of two" evolution in part (b). □

Lemma 4.5. *For any time t and any state i, if $\Pi_t(j) \ge \Sigma_t(j)$ for all states j in $[i-2, i+2]$, then $\Pi_{t+2}(i) \ge \Sigma_{t+2}(i)$.*

Proof. In the following calculations, we lean heavily on the fact that we are dealing with birth-and-death chains. Utilizing natural notation such as $K^2(h, \le i)$ for $\sum_{j \le i} K^2(h, j)$, we find using summation by parts that

$$\begin{aligned} \Pi_{t+2}(i) &= \sum_{h=0}^{i+2} \pi_t(h) K^2(h, \le i) \\ &= \sum_{j=0}^{i+2} \Pi_t(j) \left[K^2(j, \le i) - K^2(j+1, \le i)\right] \\ &= \sum_{j=i-2}^{i+2} \Pi_t(j) \left[K^2(j, \le i) - K^2(j+1, \le i)\right]. \end{aligned}$$

Recalling that K^2 is monotone, the expression in square brackets here is nonnegative, so first by hypothesis and then by reversing the above steps (now with Σ in place of Π) we have

$$\Pi_{t+2}(i) \geq \sum_{j=i-2}^{i+2} \Sigma_t(j) \left[K^2(j, \leq i) - K^2(j+1, \leq i)\right] = \sum_{h=0}^{i+2} \sigma_t(h) K^2(h, \leq i).$$

But $K_0^2 \preceq K^2$ (as noted in Section 4.2) and σ_t is non-increasing [Lemma 4.4(a)], so we finally conclude

$$\Pi_{t+2}(i) \geq \sum_{h=0}^{i+2} \sigma_t(h) K_0^2(h, \leq i) = \Sigma_{t+2}(i),$$

as desired. □

An immediate consequence is the following:

Lemma 4.6. *If $p_0 \leq 1/2$, then $\Pi_t(i) \geq \Sigma_t(i)$ for all times t and all states i.*

Proof. As previously discussed, we need only consider odd times, for which the proof is immediate by induction using Lemma 4.5 once the basis $t = 1$ is handled. But indeed

$$\Pi_1(0) = 1 - p_0 \geq \tfrac{1}{2} = \Sigma_1(0)$$

and $\Pi_1(i) = 1 = \Sigma_1(i)$ for $i \geq 1$. □

We can also prove that $\Pi_t(i) \geq \Sigma_t(i)$ for all t if the transition probability from i to $i+1$ is sufficiently low:

Lemma 4.7. *For any state i such that $p_i \leq 1/2$, we have $\Pi_t(i) \geq \Sigma_t(i)$ for all times t.*

Proof. We begin with the observation that, by last-step analysis,

$$\Pi_t(i) = \Pi_{t-1}(i-1) + \pi_{t-1}(i)(1 - p_i) + \pi_{t-1}(i+1) p_i,$$

which can be rewritten in terms of cdfs as

$$\Pi_t(i) = p_i \Pi_{t-1}(i+1) + (1 - 2p_i)\Pi_{t-1}(i) + p_i \Pi_{t-1}(i-1)$$

in general and as

$$\Sigma_t(i) = \tfrac{1}{2}\Sigma_{t-1}(i+1) + \tfrac{1}{2}\Sigma_{t-1}(i-1)$$

for the uniform chain.

Again we need only prove the lemma for odd times t, and then we find

$$\begin{aligned}
\Pi_t(i) &= p_i \Pi_{t-1}(i+1) + (1 - 2p_i)\Pi_{t-1}(i) + p_i \Pi_{t-1}(i-1) \\
&\geq p_i \Sigma_{t-1}(i+1) + (1 - 2p_i)\Sigma_{t-1}(i) + p_i \Sigma_{t-1}(i-1) \\
&\geq \tfrac{1}{2}\Sigma_{t-1}(i+1) + \tfrac{1}{2}\Sigma_{t-1}(i-1) \\
&= \Sigma_t(i),
\end{aligned}$$

where we know the first inequality holds because $t-1$ is even (whence Π_{t-1} dominates Σ_{t-1}) and $p_i \leq 1/2$, and the second inequality follows from concavity of Σ_{t-1} [Lemma 4.4(c)] again using $p_i \leq 1/2$. □

We can now combine Lemmas 4.5 and 4.7 to prove:

Lemma 4.8. *If $p_i \leq 1/2$ and $p_{i+1} \leq 1/2$, then for all times t we have*

$$\Pi_t(j) \geq \Sigma_t(j) \text{ for all } j \geq i+2. \tag{4.9}$$

Proof. We need only consider odd times, and we proceed by induction on t. For $t = 1$ we have $\Pi_1(j) = 1 = \Sigma_1(j)$ for all $j \geq 2$; so we move on to the induction step.

Suppose that (4.9) holds with t replaced by $t-2$. Use of Lemma 4.7 then ensures that we in fact have $\Pi_{t-2}(j) \geq \Sigma_{t-2}(j)$ for all $j \geq i$. Hence for any $j \geq i+2$ we have $\Pi_{t-2}(\ell) \geq \Sigma_{t-2}(\ell)$ for all $\ell \in [j-2, j+2]$ and therefore, by Lemma 4.5, $\Pi_t(j) \geq \Sigma_t(j)$. □

Lemma 4.9. *If $t + i$ is even, then*

$$\Pi_t(i) \geq \Sigma_t(i).$$

Proof. We may assume that t and i are odd. In light of Lemma 4.6, we may also assume $p_0 > 1/2$. Let 2ℓ be the first state where the alternation of p_i's greater than and no greater than $1/2$ is broken:

$$\begin{aligned} & p_{2\ell} \leq \tfrac{1}{2}, \\ \forall\, 0 \leq m < \ell: \quad & p_{2m} > \tfrac{1}{2} \text{ and } p_{2m+1} \leq \tfrac{1}{2}. \end{aligned} \tag{4.10}$$

(If there is no such break, we define 2ℓ to be $n+1$ or $n+2$ according as n is odd or even.) Notice that the break happen only at an even state, since two consecutive p_i's cannot both exceed $1/2$.

Since i is odd, we have either $i < 2\ell$ or $i > 2\ell$. In the former case, condition (4.10) implies $p_i \leq 1/2$, and Lemma 4.7 proves that $\Pi_t(i) \geq \Sigma_t(i)$. In the latter case, we must have $2\ell \leq n-1$ in order for i to be a state; we then observe that $p_{2\ell-1} \leq 1/2$ and $p_{2\ell} \leq 1/2$, and then $\Pi_t(i) \geq \Sigma_t(i)$ by Lemma 4.8. □

We are now prepared to complete the proof of Theorem 4.3.

Proof of Theorem 4.3. Because the cdf inequality (4.6) holds when either t is even or (by Lemma 4.9) when $t + i$ is even, we need only establish the asserted majorization when t is odd and i is even. Indeed, in that case using Lemma 4.4(d) we have

$$\begin{aligned} \Sigma_t(i) = \tfrac{1}{2}[\Sigma_t(i-1) + \Sigma_t(i+1)] &\leq \tfrac{1}{2}[\Pi_t(i-1) + \Pi_t(i+1)] \\ &\leq \Pi_t(i-1) + \max\{\pi_t(i), \pi_t(i+1)\}, \end{aligned}$$

and so there exist $i+1$ entries of the vector π_t whose sum is at least $\Sigma_t(i)$. We conclude that π_t majorizes σ_t, as asserted. □

Remark 4.10. (a) The multiset of values $\{\mathbf{P}_i(U_t = j) : j \in \{0, \ldots, n\}\}$ for the uniform chain U started in state i does not depend on $i \in \{0, \ldots, n\}$; therefore, the uniform chain minimizes various distances from stationarity (including all those listed in Example 3.8) not only when the starting state is 0 but in the worst case over all starting states (and indeed over all starting distributions).

To see the asserted invariance in starting state, consider simple symmetric random walk V on the cycle $\{0, \ldots, 2n+1\}$, with transition probability 1/2 in each direction between adjacent states (modulo $2n+2$). Then for every $i, j \in \{0, \ldots, n\}$ we have (by regarding states $n+1, \ldots, 2n+1$ as "mirror reflections" of the states $n, \ldots, 0$, respectively)

$$\mathbf{P}_i(U_t = j) = \mathbf{P}_i(V_t = j) + \mathbf{P}_i(V_t = 2n+1-j),$$

where at most one of the two terms on the right—namely, the one with $j - i \equiv t$ (modulo 2)—is positive. Thus, as multisets of $2n+2$ elements each, we have the equality

$$\{\mathbf{P}_i(U_t = j) : j \in \{0, \dots, n\}\} \cup \{0, \dots, 0\} = \{\mathbf{P}_i(V_t = j) : j \in \{0, \dots, 2n+1\}\},$$

where the multiset $\{0, \dots, 0\}$ on the left here has (of course) $n+1$ elements. Since the multiset on the right clearly does not depend on i, neither does $\{\mathbf{P}_i(U_t = j) : j \in \{0, \dots, n\}\}$.

(b) The SLEM (second-largest eigenvalue in modulus) is an asymptotic measure (in the worst case over starting states) of distance from stationarity. Accordingly, by remark (a), the uniform chain minimizes SLEM among all symmetric birth-and-death chains. Thus we recover the main result of [5].

5. Fastest-mixing monotone birth-and-death chains

Let n be a positive integer and consider the state space $\mathcal{X} = \{0, \dots, n\}$. Let π be a log-concave distribution on $\mathcal{X}$, and consider the class of discrete-time monotone birth-and-death chains with state space $\mathcal{X}$ and stationary distribution π, started in state 0. In this section we identify the fastest-mixing stochastically monotone chain in this class as having kernel (call it K_π) with (death, hold, birth) probabilities (q_i, r_i, p_i) given for $i \in \mathcal{X}$ by

$$q_i = \frac{\pi_{i-1}}{\pi_{i-1} + \pi_i}, \qquad r_i = \frac{\pi_i^2 - \pi_{i-1}\pi_{i+1}}{(\pi_{i-1} + \pi_i)(\pi_i + \pi_{i+1})}, \qquad p_i = \frac{\pi_{i+1}}{\pi_i + \pi_{i+1}}, \tag{5.1}$$

with $\pi_{-1} := 0$ and $\pi_{n+1} := 0$. In Section 5.1 we first find the FMMC when π is held fixed; then in Section 5.2 we show that, when π is allowed to vary, taking it to be uniform gives the slowest mixing in separation.

Throughout, we make heavy use of reversibility. Recall that any irreducible birth-and-death chain on $\mathcal{X}$ is reversible with respect to its unique stationary distribution π.

5.1. The FMMC when π is fixed.

The main result of this subsection is the following comparison inequality; and then Proposition 3.2 and Corollary 3.3 establish three senses (TV, separation, and L^2) in which the chain with kernel K_π is fastest-mixing.

Theorem 5.1. *Let π be log-concave on $\mathcal{X} = \{0, \dots, n\}$. Let K_π have (death, hold, birth) probabilities (q_i, r_i, p_i) given by* (5.1)*. Then K_π is a monotone birth-and-death kernel with stationary distribution π, and $K_\pi \preceq K$ for any such kernel K.*

Proof. Since for each i the numbers q_i, r_i, p_i are nonnegative (r_i because of the log-concavity of π) and sum to unity, K_π is indeed a birth-and-death kernel. Since $\pi_i p_i \equiv \pi_{i+1} q_{i+1}$, it is reversible with stationary distribution π. Since $p_i + q_{i+1} \equiv 1$, it satisfies the inequality (4.1) and so is monotone.

We now consider monotone birth-and-death kernels K with stationary distribution π and *general* (q_i, r_i, p_i). We prove $K_\pi \preceq K$ by extending the calculations in Section 4 and in particular in Section 4.1. Note that if f is the indicator of the down-set $\{0, 1, \dots, \ell\}$, then Kf satisfies

$$(Kf)_j = \begin{cases} 1 & \text{if } 0 \le j \le \ell - 1 \\ 1 - p_\ell & \text{if } j = \ell \\ q_{\ell+1} & \text{if } j = \ell + 1 \\ 0 & \text{otherwise;} \end{cases} \tag{5.2}$$

hence if g is the indicator of the down-set $\{0, 1, \dots, m\}$, then

$$\langle Kf, g\rangle = \begin{cases} \sum_{j=0}^{m} \pi_j & \text{if } 0 \le m \le \ell - 1 \\ \sum_{j=0}^{\ell} \pi_j - \pi_\ell p_\ell & \text{if } m = \ell \\ \sum_{j=0}^{\ell} \pi_j & \text{if } \ell + 1 \le m \le n. \end{cases} \tag{5.3}$$

Monotonicity (4.1) requires precisely that for each $\ell = 0, \dots, n-1$ we have

$$p_\ell \left(1 + \frac{\pi_\ell}{\pi_{\ell+1}}\right) = p_\ell + q_{\ell+1} \le 1,$$

so clearly $K_\pi \preceq K$. □

Remark 5.2. We see more generally that the kernels $K \in \mathcal{F}$ are non-increasing (in $\preceq$) in each p_i and that $p_i = \pi_{i+1}/(\pi_i + \pi_{i+1})$ maximizes p_i subject to the monotonicity constraint. (This remark generalizes Remark 4.1.) We observe in passing that the identity kernel I is the top element (i.e., unique maximal element) in the restriction of the comparison-inequality partial order $\preceq$ to monotone birth-and-death chains.

Example 5.3. Suppose that the stationary pmf is proportional to $\pi_i \equiv \rho^i$, i.e., is either truncated geometric (if $\rho < 1$) or its reverse (if $\rho > 1$) or uniform (if $\rho = 1$). Then the kernel K_π corresponds to biased random walk:

$$q_i \equiv q := \frac{1}{1+\rho}, \qquad r_i \equiv 0, \qquad p_i \equiv p := \frac{\rho}{1+\rho}, \tag{5.4}$$

with the endpoint exceptions, of course, that $q_0 = 0$, $r_0 = q$, $r_n = p$, $p_n = 0$.

5.2. **Slowest FMMC: the uniform chain.** In this subsection we consider the monotone FMMCs given by (5.1) for log-concave pmfs π and show (Theorem 5.9) that the uniquely slowest to mix in separation (at every time t) is obtained by setting π = uniform. Our first two results of this subsection consider ergodic birth-and-death chains and their so-called strong stationary duals and do not need any assumption about log-concavity of π. By "ergodic" we mean that the chain is assumed to be aperiodic, irreducible, and positive recurrent (the third of which follows automatically from the first two since our state space is finite) and so settles down to its unique stationary distribution.

Proposition 5.4. *Let X be an ergodic monotone birth-and-death chain on $\mathcal{X} = \{0, \dots, n\}$ with stationary pmf π, (death, hold, birth) transition probabilities (q_i, r_i, p_i) satisfying*

$$q_{i+1} + p_i = 1 \qquad (i = 0, \dots, n-1), \tag{5.5}$$

and initial state 0*. Let H denote the cdf corresponding to π, with $H_{-1} := 0$, and set*

$$q_i^* = \frac{H_{i-1}}{H_i} p_i, \quad r_i^* = 0, \quad p_i^* = \frac{H_{i+1}}{H_i} q_{i+1} \qquad (i = 0, \dots, n-1). \tag{5.6}$$

Then

$$\text{sep}(t) = \mathbf{P}(T > t) \qquad (t = 0, 1, \dots),$$

where the random variable T is the hitting time of state n for the birth-and-death chain X^* with initial state 0 *and transition probabilities* (5.6).

Proof. The chain X^* is called the strong stationary dual (SSD) of X, and the proposition is an immediate consequence of SSD theory [8, Section 4.3]. □

Example 5.5. For a biased random walk as discussed in Example 5.3, the dual kernel is

$$q^*_i = \frac{1-\rho^i}{1-\rho^{i+1}} \times \frac{\rho}{1+\rho}, \quad r^*_i = 0, \quad p^*_i = \frac{1-\rho^{i+2}}{1-\rho^{i+1}} \frac{1}{1+\rho} \qquad (i = 0, \dots, n-1).$$

It is easy to check that we obtain the *same* dual kernel for ratio ρ^{-1} as for ρ. Thus if q and p are interchanged in a biased random walk with no holding except at the endpoints, then the two chains mix equally quickly in separation.

This can be seen another way: More generally, if the state space is a partially ordered set possessing both bottom $(\hat{0})$ and top $(\hat{1})$ elements, then for any ergodic kernel K such that both K and the time-reversal $\widetilde{K}$ are stochastically monotone, *the chain K from $\hat{0}$ and the chain $\widetilde{K}$ from $\hat{1}$ mix equally quickly in separation.* Indeed, it is easy to see that for every t we have, in obvious notation,

$$\mathrm{sep}_{\hat{0}}(t) = 1 - \frac{K^t(\hat{0}, \hat{1})}{\pi_{\hat{1}}} = 1 - \frac{\widetilde{K}^t(\hat{1}, \hat{0})}{\pi_{\hat{0}}} = \widetilde{\mathrm{sep}}_{\hat{1}}(t).$$

Lemma 5.6. *Let K and L be two ergodic monotone birth-and-death chains on $\mathcal{X} = \{0, \dots, n\}$, both started at 0, with possibly different stationary distributions. Suppose that $K(i+1, i) + K(i, i+1) = 1 = L(i+1, i) + L(i, i+1)$. Consider the notation of* (5.6) *and suppose also that p^*_i arising from Y is at least p^*_i arising from Z for all $i = 0, \dots, n$. Then Y mixes faster in separation*[3] *than does Z.*

Proof. Let Y^* and Z^* be the corresponding SSDs, as in Proposition 5.4. An obvious coupling gives $Y^*_t \ge Z^*_t$ for every t, and the lemma follows. It is worth pointing out that while the dual chains may not be monotone, this causes no problem with the coupling because Y^*_t and Z^*_t must have the same parity for every t; that's because the holding probabilities for both dual chains all vanish. □

Next, given a FMMC for log-concave π, we show that it mixes faster in separation than does a certain biased random walk.

Theorem 5.7. *Consider the fastest-mixing monotone birth-and-death chain X with log-concave stationary pmf π, kernel* (5.1)*, and initial state 0. Define*

$$\rho_i := \pi_{i+1}/\pi_i \quad (i = 0, \dots, n-1),$$

and suppose that $i = i_0$ minimizes $|\ln \rho_i|$. Then X mixes faster in separation than does the biased random walk (5.4) *with ρ set to ρ_{i_0}.*

Proof. Log-concavity is precisely the condition that ρ_k is non-increasing in k. Hence p^*_i satisfies

$$\begin{aligned} p^*_i &= \frac{H_{i+1}}{H_i} \frac{\pi_i}{\pi_i + \pi_{i+1}} \\ &= \left(1 + \rho_i \frac{\pi_i}{H_i}\right) \times \frac{1}{1+\rho_i} = \left(1 + \frac{\rho_i}{\sum_{j=0}^{i} \prod_{k=j}^{i-1} \rho_k^{-1}}\right) \times \frac{1}{1+\rho_i} \\ &\ge \left(1 + \frac{\rho_i}{\sum_{j=0}^{i} \rho_i^{-(i-j)}}\right) \times \frac{1}{1+\rho_i} = f_i(\rho_i), \end{aligned} \tag{5.7}$$

[3] Recall our terminological convention stated in the paragraph preceding Corollary 3.3.

where the function

$$f_i(\rho) := \frac{1-\rho^{i+2}}{1-\rho^{i+1}}\frac{1}{1+\rho} \qquad \left[\text{with } f_i(1) := \frac{i+2}{2(i+1)}\right] \tag{5.8}$$

satisfies $f_i(\rho^{-1}) \equiv f_i(\rho)$ and can be shown by induction on i to be non-increasing in $\rho \le 1$ (and strictly so for $i \ge 1$). The induction step uses the fact that

$$f_i(\rho) = 1 - \frac{\rho}{(1+\rho)^2 f_{i-1}(\rho)}$$

together with the induction hypothesis and the (strict) increasingness of the function $\rho \mapsto \rho/(1+\rho)^2$ for $\rho \le 1$. Therefore

$$p_i^* \ge f_i(\rho_{i_0}),$$

and this last expression is the dual birth probability from state i for the biased random walk with ratio ρ_{i_0}. The conclusion of the theorem now follows from Lemma 5.6. □

So the question as to which of the FMMCs (5.1) is slowest to mix is reduced to finding the slowest biased random walk. But we've already done the calculations needed to prove the following result:

Theorem 5.8. *Consider biased random walks as in Example 5.3, each with initial state* 0*. The walks are monotonically slower to mix in separation as* $\min\{p/q, q/p\}$ *increases.*

Proof. We have already noted at Example 5.5 that the speed of mixing is invariant under interchange of p and q. Moreover, as $\rho = p/q$ increases over $(0, 1]$, the chains are monotonically slower to mix in separation because we have equality in (5.7) and hence

$$p_i^* = f_i(\rho),$$

which (as shown in the proof of Theorem 5.7) is non-increasing in $\rho \le 1$. □

The next theorem is the main result of the subsection and is an immediate corollary of Theorems 5.7 and 5.8.

Theorem 5.9. *Among the fastest-mixing monotone birth-and-death chains* (5.1) *with initial state* 0 *and log-concave stationary pmf* π*, the uniform chain is slowest to mix in separation.*

Remark 5.10. How fast *does* an ergodic monotone birth-and-death chain mix in separation? We have addressed this question in general in Proposition 5.4 and in the last sentence of Example 5.5. The biased random walk (5.4) is treated in some detail in [11, Section XVI.3]. We note:

(a) The eigenvalues, listed in decreasing order, are 1 and

$$2\sqrt{pq}\cos\frac{\pi j}{n+1} \qquad (j = 1, \dots, n).$$

(b) Fix ρ and consider $n \to \infty$. Let $\mu = |p - q|$ denote the size of the drift of the walk. If $\mu \ne 0$ (i.e., $\rho \ne 1$), there is a "cutoff phenomenon" for separation at time $t = \mu n + c_\rho n^{1/2}$. This means (roughly put) that separation is small at that time t when c_ρ is near $-\infty$ and large when it is near $+\infty$, with the subscript in c_ρ indicating that the definition of "near" depends on ρ.

(c) If $\rho = 1$ (the uniform chain), it takes time of the larger order n^2 for separation to drop from near 1 to near 0, and in this case there is no cutoff phenomenon.

6. Lovász–Winkler mixing times

In previous sections we have discussed mixing in terms of TV, separation, L^2, and other functions measuring discrepancy. An alternative description of speed of convergence is provided by mixing times as defined by Lovász and Winkler [20]; according to their definition (reviewed below), and unlike for our previous notions of mixing, one number ["the mixing time", $T_{\text{mix}}(X)$] is assigned to each chain X.

In this section we compute $T_{\text{mix}}(X)$ for any irreducible birth-and-death chain X started at 0 and then revisit the FMMC problems of the preceding two sections using T_{mix} as our criterion. One highlight is this: For the path-problem on $\mathcal{X} = \{0, \dots, n\}$, we show that the uniform chain is the fastest-mixing symmetric birth-and-death chain in the sense of Lovász and Winkler [20] if and only if n is even, and we identify the fastest chain when n is odd.

According to the definition in [20], the *mixing time* for any irreducible (discrete-time) finite-state Markov chain X having stationary distribution π is the (attained) infimum of expectations of randomized stopping times for which π is the distribution of the stopping state. In symbols,

$$T_{\text{mix}}(X) := \min \mathbf{E}\, S \tag{6.1}$$

where the infimum is taken over randomized stopping times S such that the distribution of X_S is π. For computing $T_{\text{mix}}(X)$, a very useful theorem from [20] asserts that a randomized stopping time S achieves the minimum in (6.1) if and only if it has a halting state, that is, a state x such that if $X_t = x$ then (almost surely) $S \le t$. We will use this result to compute $T_{\text{mix}}(X)$ for any irreducible birth-and-death chain in Theorem 6.2, but first we state a lemma about expected hitting times for birth-and-death chains.

Lemma 6.1. *For an irreducible birth-and-death chain on $\mathcal{X} = \{0, \dots, n\}$ (in discrete or continuous time) with stationary distribution π and initial state* 0*, let T denote the hitting time of state n.*

(a) *In discrete time, denote the birth probability from state i by p_i. Then*

$$\mathbf{E}\, T = \sum_{i=0}^{n-1} \frac{1}{\pi_i p_i} \sum_{k=0}^{i} \pi_k.$$

(b) *In continuous time, denote the birth rate from state i by λ_i. Then*

$$\mathbf{E}\, T = \sum_{i=0}^{n-1} \frac{1}{\pi_i \lambda_i} \sum_{k=0}^{i} \pi_k.$$

Proof. Each assertion is easily established, and each follows immediately from the other; for (b), see, e.g., [18, Chapter 4, Problem 22]. □

Theorem 6.2. *Let X be an irreducible (discrete-time) birth-and-death chain on $\mathcal{X} = \{0, \dots, n\}$ with stationary pmf (respectively, cdf) π (resp., H) and initial state* 0*. Then*

$$T_{\text{mix}}(X) = \sum_{i=0}^{n-1} \frac{H_i(1 - H_i)}{\pi_i p_i}.$$

Proof. Let us use the *naive rule* S as our randomized stopping time: Choose j randomly according to π, and then let S be the hitting time of j. Obviously the stopping distribution is π, as required. Moreover, the state j must be hit en route to n; hence n is a halting state and S achieves the minimum at (6.1).

To compute $T_{\text{mix}}(X) = \mathbf{E}\, S$, we first note that Lemma 6.1(a) yields (easily) corresponding formulas for the expected value of the hitting time T_j of each state j:

$$\mathbf{E}\, T_j = \sum_{i=0}^{j-1} \frac{H_i}{\pi_i p_i}.$$

Therefore

$$T_{\text{mix}}(X) = \sum_{j=0}^{n} \pi_j \mathbf{E}\, T_j = \sum_{j=0}^{n} \pi_j \sum_{i=0}^{j-1} \frac{H_i}{\pi_i p_i} = \sum_{i=0}^{n-1} \frac{H_i(1 - H_i)}{\pi_i p_i},$$

as desired. □

Remark 6.3. (a) The Lovász–Winkler theory of mixing times and the statement and proof of Theorem 6.2 all carry over routinely to the "continuized" chain which evolves in the same way as the given discrete-time chain but with independent exponential random times with mean 1 replacing unit times. In particular, the value of $T_{\text{mix}}(X)$ remains unchanged under continuization of an irreducible discrete-time birth-and-death chain X with initial state 0.

(b) By a theorem of Aldous and Diaconis [2, Proposition 3.2] in discrete time and a theorem of Fill [14, Theorem 1.1] in continuous time, any ergodic finite-state Markov chain X (regardless of initial distribution) has a fastest (i.e., stochastically minimal) strong stationary time T satisfying $\mathbf{P}(T > t) = \text{sep}(t)$ for every t (restricted to integer values for a discrete-time chain). If the state space is partially ordered with bottom element $\hat{0}$ and top element $\hat{1}$ and the chain X starts in $\hat{0}$, and if the time-reversed kernel $\widetilde{K}$ is monotone, then $\hat{1}$ is a halting state for any such T; to see this, observe that

$$\begin{aligned}\mathbf{P}(X_t = \hat{1},\, T > t) &= \mathbf{P}(X_t = \hat{1}) - \mathbf{P}(T \le t,\, X_t = \hat{1}) \\ &= \pi_{\hat{1}} \left[\frac{K^t(\hat{0}, \hat{1})}{\pi_{\hat{1}}} - (1 - \text{sep}(t)) \right] \\ &= \pi_{\hat{1}} \left[\min_i \frac{K^t(\hat{0}, i)}{\pi_i} - (1 - \text{sep}(t)) \right] = 0,\end{aligned}$$

where π is the stationary distribution and the penultimate equality follows from the monotonicity of $\widetilde{K}^t$.

Now consider an ergodic birth-and-death chain X (in discrete or continuous time) on $\mathcal{X} = \{0, \dots, n\}$ with stationary distribution π and initial state 0. In the discrete-time case, assume that the chain is monotone; this is automatic in continuous time by a simple and standard coupling argument. Then a fastest (i.e., stochastically minimal) strong stationary time T exists, and n is a halting state for any such T. It follows that $T_{\text{mix}}(X) = \mathbf{E}\, T$ and thus Theorem 6.2 also gives an expression for $\mathbf{E}\, T$, which equals

$$\sum_{t=0}^{\infty} \mathbf{P}(T > t) = \sum_{t=0}^{\infty} \text{sep}(t)$$

in discrete time and equals

$$\int_0^\infty \mathbf{P}(T > t)\, dt = \int_0^\infty \operatorname{sep}(t)\, dt$$

in continuous time. This remark gives added import to the value of $T_{\mathrm{mix}}(X)$ for any irreducible discrete-time birth-and-death chain X (whether monotone or not) with initial state 0: It equals the integral of separation for the continuized chain.

(c) Given a collection $\mathcal{C}$ of irreducible discrete-time birth-and-death chains Y with initial state 0, suppose that $X \in \mathcal{C}$ satisfies $X = \arg\min_{Y \in \mathcal{C}} T_{\mathrm{mix}}(Y)$. In light of remark (b), one might wonder whether the continuized chain corresponding to X minimizes $\operatorname{sep}(t)$ at *every* time t over all continuizations of chains $Y \in \mathcal{C}$. Theorem 6.5(b) provides a counterexample. Indeed, it can be shown that if we compare the chain of the form (6.4) but with θ_n changed to $(n-1)/(2n)$ with any other birth-and-death chain having initial state 0 and symmetric kernel K, then there exists $t_0 = t_0(K)$ such that continuized separation at time t is strictly smaller for the former chain than for the latter for all $0 < t \le t_0$.[4] Likewise, in the "ladder game" discussed in Section 7 it is the uniform chain, not the chain discussed there, that is "best in separation for small t" in similar fashion.

We are now in position to determine, for given π, the birth-and-death chain X that minimizes $T_{\mathrm{mix}}(X)$ among those having initial state 0, stationary distribution π, and no holding probability except at the endpoints of the state space. Unlike in Section 5, we do not need to restrict to monotone kernels; and rather than assuming that π is log-concave, we assume instead that π is non-decreasing. For the case that π is uniform, we will give later an argument that removes the restriction about holding probabilities. [There are examples, such as $\pi = \frac{1}{15}(1, 2, 4, 4, 4)$, showing that the restriction cannot be removed in general.]

Theorem 6.4. *Let $\mathcal{X} = \{0, \dots, n\}$. Among all irreducible birth-and-death chains X having a given positive non-decreasing stationary pmf π, initial state 0, and no holding probability except at 0 and n, there is a unique chain X_π minimizing $T_{\mathrm{mix}}(X)$. Moreover,*

(a) *Let $a_i := \sum_{j=1}^{i} (-1)^{i-j} \pi_j$ for $i = 0, \dots, n-1$. Define*

$$f(w) := \sum_{i=0}^{n-1} \frac{H_i(1 - H_i)}{(-1)^i w + a_i}.$$

Then there exists a unique w_π minimizing $f(w)$ over $w \in [0, \pi_0]$, and $T_{\mathrm{mix}}(X_\pi) = f(w_\pi)$.

(b) *The optimal chain X_π has transition probabilities*

$$q_i = \frac{a_{i-1} + (-1)^{i-1} w_\pi}{\pi_i}, \qquad r_i = 0, \qquad p_i = \frac{a_i + (-1)^i w_\pi}{\pi_i} \qquad (i = 0, \dots, n)$$

with the exceptions $q_0 = 0$, $r_0 = 1 - p_0$, $r_n = 1 - q_n$, and $p_n = 0$.

[4] Indeed, if Y and Z are the discrete-time and continuized chain corresponding to K, then, with π denoting the uniform pmf, as $t \to 0$ we find

$$1 - \operatorname{sep}_Z(t) = \frac{\mathbf{P}(Z_t = n)}{\pi_n} = e^{-t} \frac{t^n}{n!} \frac{\mathbf{P}(Y_n = n)}{\pi_n} + o(t^{n+1}) = \frac{t^n}{n!} (n+1) p_0 p_1 \dots p_{n-1} + o(t^{n+1}),$$

and $p_0 p_1 \dots p_{n-1}$ is uniquely maximized subject to $p_{k-1} + p_k \le 1$ for $k = 0, \dots, n-1$ by choosing $p_k = (n+1)/(2n)$ if k is even and $p_k = (n-1)/(2n)$ if k is odd.

Proof. We begin by noting that birth-and-death kernels with stationary distribution π (in complete generality, irrespective of holding probabilities or non-decreasingness of π) are in one-to-one correspondence with nonnegative sequences $\mathbf{w} = (w_{-1}, w_0, \ldots, w_n)$ satisfying $w_{-1} = 0 = w_n$ and

$$w_{i-1} + w_i \leq \pi_i \qquad (i = 0, \ldots, n), \tag{6.2}$$

the correspondence being $w_i = \pi_i p_i = \pi_{i+1} q_{i+1}$, $i = 0, \ldots, n-1$. The proof is easy, and the correspondence gives

$$r_i = 1 - q_i - p_i = 1 - \frac{w_{i-1} + w_i}{\pi_i} \qquad (i = 0, \ldots, n)$$

for the holding probabilities. In this $\mathbf{w}$-parameterization, Theorem 6.2 gives

$$T_{\text{mix}} = \sum_{i=0}^{n-1} \frac{H_i(1 - H_i)}{w_i}. \tag{6.3}$$

The constraint $r_i = 0$ for $i = 0, \ldots, n-1$ is precisely the constraint that equality holds in (6.2) for $i = 1, \ldots, n-1$. Then we must have $w := w_0 \in [0, \pi_0]$ and

$$w_i = (-1)^i w + a_i \qquad (i = 0, \ldots, n-1).$$

It follows from the assumption that π is non-decreasing that these w_i's are indeed all nonnegative [and all positive if $w \in (0, \pi_0)$]. This proves the theorem, because f is continuous on $[0, \pi_0]$ and both finite and strictly convex[5] on $(0, \pi_0)$. □

We now specialize to the case of uniform π, removing the restriction on holding from Theorem 6.4 and solving explicitly for the value w in Theorem 6.4(a). We find it somewhat surprising that the chain minimizing T_{mix} is *not* the uniform chain whenever $n \geq 3$ is odd.

Theorem 6.5. *Consider the problem of minimizing T_{mix} among all birth-and-death chains on $\mathcal{X} = \{0, \ldots, n\}$ with initial state 0 and symmetric kernel.*

(a) *If $n \geq 2$ is even, then the uniform chain is the unique minimizing chain.*

(b) *If n is odd, then*

$$p_k = \begin{cases} 1 - \theta_n & \text{if } k \text{ is even} \\ \theta_n & \text{if } k \text{ is odd} \end{cases} \qquad (k = 0, \ldots, n-1) \tag{6.4}$$

gives the unique minimizing chain, where for any m we define

$$\theta_{m-1} := \frac{1}{6}\left[\sqrt{(m^2+2)(m^2-4)} - (m^2 - 4)\right]. \tag{6.5}$$

We have written the formula for θ_{m-1} rather than that for θ_n because it is simpler to write.

Remark 6.6. Although the uniform chain is not optimal when n is odd, it is nearly optimal, since θ_n has the asymptotics

$$\theta_n = \tfrac{1}{2} - \tfrac{3}{4} n^{-2} + O(n^{-3}) \quad \text{as } n \to \infty$$

[5] In the general setting of (6.3), T_{mix} is a strictly convex function on a nonempty convex domain (an intersection of half-spaces) of arguments $\mathbf{w}$ and so has a unique minimum. The optimal $\mathbf{w}$ is on the boundary of the domain; more specifically, for every $i = 0, \ldots, n-1$, if the optimal $\mathbf{w}$ does not lie on the hyperplane delimiting the ith half-space (6.2), then it lies on the $(i+1)$st such hyperplane.

and the value of T_{mix} (recall Theorem 6.2) for $p_k \equiv 1/2$ is $\frac{1}{3}n^2 + n + \frac{2}{3}$, only slightly larger than the optimal value $\frac{1}{3}n^2 + n + \frac{2}{3} - \frac{3}{4}n^{-2} + O(n^{-3})$.

Proof of Theorem 6.5. Recall Theorem 6.2; thus the goal is to minimize

$$f(\mathbf{p}) := \sum_{k=0}^{n-1} \frac{(k+1)(n-k)}{p_k}$$

over vectors $\mathbf{p} = (p_0, \ldots, p_{n-1})$ that are nonnegative (we won't repeat this nonnegativity condition below) and satisfy

$$p_{k-1} + p_k \leq 1 \text{ for } k = 0, \ldots, n \tag{6.6}$$

where $p_{-1} = 0 = p_n$. The objective function $f(\mathbf{p})$ is strictly convex in $\mathbf{p}$ (by strict convexity of $x \mapsto x^{-1}$). Hence there is a unique minimizer, and because $(p_{n-1}, \ldots, p_0)$ is clearly a minimizer if $(p_0, \ldots, p_{n-1})$ is, the unique minimizer is of the form

$$(p_0, \ldots, p_{(n/2)-1}, p_{(n/2)-1}, \ldots, p_0)$$

if n is even and of the form

$$(p_0, \ldots, p_{(n-3)/2}, p_{(n-1)/2}, p_{(n-3)/2}, \ldots, p_0)$$

if n is odd. We now break into the two cases.

(a) For n even, we seek equivalently to minimize

$$f(\mathbf{p}) = 2 \sum_{k=0}^{(n/2)-1} \frac{(k+1)(n-k)}{p_k}$$

subject to

$$p_{k-1} + p_k \leq 1 \quad \text{for } k = 0, \ldots, (n/2).$$

[Note that the last of these conditions is $p_{(n/2)-1} \leq 1/2$.]

We claim (by induction on K) for $1 \leq K \leq (n/2) - 1$ that the minimizer of $\sum_{k=0}^{K} \frac{(k+1)(n-k)}{p_k}$ subject to (nonnegativity and) $p_{k-1} + p_k \leq 1$ for $k = 0, \ldots, K$ and $p_K \leq 1/2$ is $p_k \equiv 1/2$.

For the basis $K = 1$ of the induction, we seek to minimize

$$\frac{n}{p_0} + \frac{2(n-1)}{p_1}$$

subject to $p_0 + p_1 \leq 1$ and $p_1 \leq 1/2$. Clearly we should take $p_0 = 1 - p_1$ (regardless of p_1), and then we need to minimize

$$\frac{n}{1 - p_1} + \frac{2(n-1)}{p_1}$$

subject to $p_1 \leq 1/2$. Because $2(n-1) \geq n$ (i.e., $n \geq 2$), the minimizer is $p_1 = 1/2$ (and then $p_0 = 1/2$).

We now proceed to the induction step to move from $K-1$ to K. To minimize, clearly we should take $p_K = \min\{1/2, 1 - p_{K-1}\}$. The remainder of the proof for n even then breaks into two cases.

Case 1. If $p_{K-1} \geq 1/2$, then we take $p_K = 1 - p_{K-1}$ and our goal is to minimize

$$\sum_{k=0}^{K-2} \frac{(k+1)(n-k)}{p_k} + \frac{K(n-(K-1))}{p_{K-1}} + \frac{(K+1)(n-K)}{1 - p_{K-1}}$$

subject to $p_{k-1}+p_k \le 1$ for $0 \le k \le K-1$ and (because this is Case 1) $p_{K-1} \ge 1/2$. Because $(K+1)(n-K) \ge K(n-(K-1))$ and we have the restriction $p_{K-1} \ge 1/2$, we should set p_{K-1} as small as possible, namely, $p_{K-1} = 1/2$, and then we seek to minimize

$$\sum_{k=0}^{K-2} \frac{(k+1)(n-k)}{p_k} + \frac{K(n-(K-1))}{p_{K-1}} + \frac{(K+1)(n-K)}{1/2}$$

subject to $p_{k-1}+p_k \le 1$ for $0 \le k \le K-1$ and $p_{K-1} = 1/2$. Clearly the minimum value here is at least as large as the minimum value if we relax the last constraint to $p_{K-1} \le 1/2$. But then by induction the minimum value is achieved by setting $p_k \equiv 1/2$. This completes the proof in Case 1.

Case 2. If $p_{K-1} \le 1/2$, then we set $p_K = 1/2$ and the goal is to minimize

$$\sum_{k=0}^{K-1} \frac{(k+1)(n-k)}{p_k} + \frac{(K+1)(n-K)}{1/2}$$

subject to $p_{k-1}+p_k \le 1$ for $0 \le k \le K$ and $p_{K-1} \le 1/2$. But then again by induction the minimum value is achieved by setting $p_k \equiv 1/2$. This completes the proof in Case 2, and thereby completes the proof of part (a).

(b) For n odd, suppose without loss of generality that $n \ge 3$. We first prove that the optimum is again attained for a chain that satisfies equality in condition (6.6) at interior points k of the state space:

$$p_{k-1}+p_k = 1 \quad \text{for } k = 1, \ldots, n-1. \tag{6.7}$$

Recall that the minimizing $\mathbf{p}$ is unique and symmetric. Hence, considering the holding probability $r_k := 1 - p_{k-1} - p_k$ at state k, it suffices to show that there is an optimizing chain with $r_k = 0$ for $1 \le k \le (n-1)/2$.

We proceed by contradiction. We show that there exists $\mathbf{p}'$ satisfying (6.6) and $f(\mathbf{p}') < f(\mathbf{p})$ in each of the following three cases which, allowing arbitrary $k \in \{1, \ldots, (n-1)/2\}$, exhaust all possibilities where $r_k > 0$ for some $1 \le k \le (n-1)/2$:

(i) $r_k > 0$ and $r_{k-1} > 0$;
(ii) $r_k > 0$ and $r_{k-1} = 0$ and $p_k \ge 1/2$;
(iii) $p_k < 1/2$, and k is the largest value j in $\{1, \ldots, (n-1)/2\}$ such that $r_j > 0$.

In case (i), let

$$p'_{k-1} := p_{k-1} + \min\{r_{k-1}, r_k\}$$

and $p'_j := p_j$ otherwise.

In case (ii), first note that $k \ge 2$; indeed, were we to have $k = 1$, then (by our assumption) $r_0 = 0$ and so $p_0 = 1$; but then $p_1 = 0$, and such a $\mathbf{p}$ clearly doesn't minimize $f(\mathbf{p})$. Next, because $p_k \ge 1/2$ we must have $p_{k-1} < 1/2$ (because $r_k > 0$) and thus $p_{k-2} > 1/2$ (because $r_{k-1} = 0$). We can then let

$$p'_{k-1} := p_{k-1} + \epsilon, \quad p'_{k-2} := p_{k-2} - \epsilon$$

for suitably small $\epsilon > 0$, and $p'_j := p_j$ otherwise. Since $k \le (n-1)/2$, we know $k(n+1-k) > (k-1)(n+2-k)$, so the derivative of $f(\mathbf{p})$ in the direction of the vector $\delta_{k-1} - \delta_{k-2}$ is negative and $f(\mathbf{p}') < f(\mathbf{p})$.

In case (iii) we have $p_{k+2i} = p_k$ for $0 \le i \le \frac{n-1}{2} - k$, and $p_{k+2i-1} = 1 - p_k$ for $1 \le i \le \frac{n-1}{2} - k$. We form $\mathbf{p}'$ by changing these values to $p'_{k+2i} := p_k + \epsilon$ and $p'_{k+2i-1} := 1 - p_k - \epsilon$ for suitably small $\epsilon > 0$ and setting $p'_j := p_j$ otherwise. We see that $f(\mathbf{p}') < f(\mathbf{p})$ if the derivative with respect to p_k of the following expression is negative for all $p_k < 1/2$:

$$\frac{1}{p_k} \sum_{i=0}^{\frac{n-1}{2}-k} (k+2i+1)(n-k-2i) + \frac{1}{1-p_k} \sum_{i=1}^{\frac{n-1}{2}-k} (k+2i)(n+1-k-2i);$$

and that is true if (and only if) the first sum is at least as large as the second. Indeed, the first sum *is* larger than the second:

$$\begin{aligned}\sum_{i=0}^{\frac{n-1}{2}-k} (k+2i+1)(n-k-2i) - \sum_{i=1}^{\frac{n-1}{2}-k} (k+2i)(n+1-k-2i)& \\ = (k+1)(n-k) + \sum_{i=1}^{\frac{n-1}{2}-k} (n-2k-4i)& \\ = k(n-k) + \tfrac{1}{2}(n+1) > 0.&\end{aligned}$$

Since we have established constraint (6.7), *every* feasible vector $\mathbf{p}$ is of the form

$$p_k \equiv \begin{cases} 1-\theta & \text{if } k \text{ is even} \\ \theta & \text{if } k \text{ is odd,} \end{cases}$$

so we need only verify that the choice $\theta = \theta_n$ as defined at (6.5) is optimal. Indeed, writing $r = (n-1)/2$ we have

$$\begin{aligned} a_n &:= \sum_{0 \le j \le r} (2j+1)(n-2j) = \frac{1}{12}(n+1)(n^2+2n+3) \\ b_n &:= \sum_{1 \le j \le r} (2j)(n-2j+1) = \frac{1}{12}(n+1)(n-1)(n+3), \end{aligned}$$

and then the optimal choice of θ, minimizing $\frac{a_n}{1-\theta} + \frac{b_n}{\theta}$, is θ_n given by

$$\theta_n = \left(1 + \sqrt{a_n/b_n}\right)^{-1}.$$

After a little bit of computation, we find that θ_n is given in accordance with equation (6.5). □

7. A “ladder” game

In this section we discuss a simple “ladder” game, where the class of kernels considered is a certain subclass of the symmetric birth-and-death kernels considered in Section 4. Our treatment involves finding the kernel that minimizes the Lovász–Winkler mixing time T_{mix}. This particular kernel is *not* one that had previously been considered as a candidate for “fastest”.

Lange and Miller [19] discusses a “ladder” game and several contexts, including an old Japanese scheme for choosing a spouse’s Christmas gift from a list of desired items, in which it arises. We refer the reader to [19] for details. A class of Markov chains that arise in modeling the ladder game (see “Model One” in [19, Section 5]) have the permutation group on $\{0, \dots, n\}$ as state space and moves that transpose items in adjacent positions; write p_i for the probability that the positions chosen are i and $i+1$, so that

$$p_0 + p_1 + \cdots + p_{n-1} = 1. \tag{7.1}$$

We will refer to (7.1) as the "ladder condition". If we follow the movement of only a single item (this is "Model Two: The path of a single marcher as a random walk among the columns of the ladder" in [19, Section 7, esp. Figure 9]), then we have precisely the class of symmetric birth-and-death kernels considered in our path-problem of Section 4, but now subject to the ladder condition. From [19, Section 8: How many rungs is enough?] we have the following quote (with notation adjusted slightly to match that of Section 4):

> We suspect (but have not shown) that for any n, the rate of convergence is maximized when rung placement is uniform. That is, the absolute value of the largest small eigenvalue is minimized when $p_i = 1/n$ for $i = 0, 1, \dots, n-1$.

(Here "largest small eigenvalue" means the eigenvalue of the kernel with largest absolute value strictly less than 1—what is called "SLEM" in [6, 5, 4].) The authors of [19] base their suspicion on calculations for $n = 2$, for which their conjecture is indeed true.

The corresponding continuous-time problem has been studied by Fielder [12] and, in a somewhat more general setting, by Sun et al. in [32, Example 5.2]. The result is that, among all continuous-time symmetric birth-and-death chains on $\{0, \dots, n\}$, started from 0, with birth rates p_i satisfying the ladder condition (7.1), the one which is fastest-mixing in the sense of minimizing relaxation time has p_i proportional to $(i+1)(n-i)$. It can be shown that these weights also uniquely minimize SLEM in discrete time, so the conjecture in [19] is false for every $n \geq 3$.[6]

One might now suspect that these parabolic weights provide a FMMC (subject to the ladder condition) in a variety of senses, at least for chains (as henceforth assumed) starting in state 0. However, working in discrete time, it is clear (a) from reviewing the discussion in Section 4.1 that there is no bottom element with respect to $\preceq$ for monotone chains satisfying the ladder condition and (b) from Remark 4.2 that there is no bottom element in $\preceq$ for squares of ladder-condition birth-and-death kernels. Further, it can be shown, switching to continuous time to match the setting of [32] and in order to bring standard techniques to bear (it is well known that all birth and death chains in continuous time are monotone), that there is no ladder-condition birth-and-death chain minimizing separation at every time. Theorem 7.1 implies that the integral of separation over all times is minimized by weights p_i proportional to the square roots $\sqrt{(i+1)(n-i)}$ of the weights minimizing SLEM.

Theorem 7.1. *For each discrete-time symmetric birth-and-death chain with state space $\{0, \dots, n\}$, initial state 0, and birth probabilities $\mathbf{p} = (p_i)$ satisfying the ladder condition (7.1), let $f(\mathbf{p})$ denote its Lovász–Winkler mixing time T_{mix}. Then the uniquely optimal (i.e., minimizing) choice of $\mathbf{p}$ is to take p_i proportional to $\sqrt{(i+1)(n-i)}$.*

Theorem 7.1 is an immediate consequence of the following corollary to the proof of Theorem 6.4, taking π to be uniform and c to be $1/n$.

[6] At the end of their Section 8, the authors of [19] also wonder, based on results for $n = 2$, whether it might be the case for all n that, except for multiplicities, the eigenvalues are the same for the permutation chain as for the single-marcher chain. This is seen to be false by the discussion in [7, Section 1.4]. But the main theorem of [7] does establish that the second-largest eigenvalues of the two chains agree.

Corollary 7.2. *Over all discrete-time birth-and-death chains on* $\{0, \ldots, n\}$ *(started at* 0*) with given stationary distribution* π *(having cdf* H*) and*

$$\sum_{k=0}^{n-1} \pi_k p_k = c \in (0, \min_i \pi_i],$$

the mixing time T_{mix} *of the chain is minimized by the choice*

$$q_k \equiv \frac{c\sqrt{H_{k-1}(1-H_{k-1})}}{\pi_k \sum_j \sqrt{H_j(1-H_j)}}, \quad p_k \equiv \frac{c\sqrt{H_k(1-H_k)}}{\pi_k \sum_j \sqrt{H_j(1-H_j)}}, \quad r_k \equiv 1 - q_k - p_k,$$

and the minimized value is

$$T_{\text{mix}} = c^{-1} \left[\sum_{k=0}^{n-1} \sqrt{H_k(1-H_k)} \right]^2 .$$

Proof. As demonstrated in the proof of Theorem 6.4, the goal is to minimize

$$T_{\text{mix}} = \sum_{i=0}^{n-1} \frac{H_i(1-H_i)}{w_i}$$

over nonnegative sequences $(w_{-1}, w_0, \ldots, w_n)$ satisfying $w_{-1} = 0 = w_n$ and

$$w_{i-1} + w_i \leq \pi_i \qquad (i = 0, \ldots, n) \tag{7.2}$$

and $\sum_{k=0}^{n-1} w_k = c$. Ignoring the constraint (7.2), the optimal choice of the weights w_i is clear, namely, $w_i \equiv \pi_i p_i$ with p_i as asserted in the statement of the theorem. But then (7.2) is automatically satisfied because we assume $c \in (0, \min_i \pi_i]$. Evaluation of the objective function at the optimizing kernel gives the optimized value of T_{mix}. □

Remark 7.3. Let $n \to \infty$. For the optimal kernel of Theorem 7.1 we have $T_{\text{mix}} \sim \frac{\pi^2}{64} n^3$, whereas for both $p_i \equiv 1/n$ (the guess for optimality in [19]) and the choice $p_i \propto (i+1)(n-i)$ minimizing SLEM we have $T_{\text{mix}} = \frac{1}{6} n(n+1)(n+2) \sim \frac{1}{6} n^3$.

8. Can extra updates delay mixing? (No, subject to positive correlations)

Can extra updates delay mixing? This question is the title of a paper [23] by Yuval Peres and Peter Winkler (see also Holroyd [17] for counterexamples). Peres and Winkler show that the answer is no, for total variation distance, in the setting of monotone spin systems, generalized by replacing the set of spins $\{0, 1\}$ by any linearly ordered set. (We review relevant terminology below.) In Theorem 8.3 we recapture and extend their result using comparison inequalities by showing that $K_v \preceq I$ for any kernel K_v that updates a single site v, i.e., that the identity kernel [as for the monotone birth-and-death example, see Remark 5.2(a)] only slows mixing (when the initial pmf has non-increasing ratio with respect to the stationary pmf)—because then, noting reversibility and stochastic monotonicity of each K_v and applying Proposition 2.4, for any $v_1, \ldots, v_t$ the product $K_{v_1} \cdots K_{v_t}$ increases in $\preceq$ by deletion of any K_{v_i}. The comparison inequality $K_v \preceq I$ holds in the more general setting of a partially ordered set of "spins", subject to the following restriction: Starting with distribution π and a site v and conditioning on the spins at all sites other than v, the conditional law of the spin at v should have positive correlations (as, of course, does any distribution on a *linearly* ordered set).

8.1. **Positive correlations.** Recall that a pmf π on a finite partially ordered set $\mathcal{X}$ is said to *have positive correlations* if (in the notation of Section 2)

$$\langle f, g \rangle \geq \langle f, 1 \rangle \langle g, 1 \rangle$$

for every $f, g \in \mathcal{M}$, and that if S is *linearly* ordered then (by "Chebyshev's other inequality"; see, e.g., [22, Lemma 16.2]) *all* probability measures have positive correlations. The connection with comparison inequalities is the following simple lemma, in relation to which we note that both K_π and I are stochastically monotone kernels possessing stationary distribution π.

Lemma 8.1. *A pmf π on a finite partially ordered set $\mathcal{X}$ has positive correlations if and only if $K_\pi \preceq I$, where K_π is the trivial kernel that jumps in one step to π and I is the identity kernel.*

Proof. Since for any f and g we have

$$\langle K_\pi f, g \rangle = \langle \langle f, 1 \rangle, g \rangle = \langle f, 1 \rangle \langle g, 1 \rangle$$

and $\langle I f, g \rangle = \langle f, g \rangle$, the lemma is proved. □

Proposition 8.2. *Let π be a pmf on a finite partially ordered set. Partition $\mathcal{X}$, suppose that a given kernel K on $\mathcal{X}$ is a direct sum [as in Proposition 2.3(c)] of trivial kernels K_i (as in Lemma 8.1) on the cells of the partition, and suppose that π conditioned to each cell has positive correlations. Then $K \preceq I$.*

Proof. Simply combine Lemma 8.1 and Proposition 2.3(c). □

8.2. **Monotone spin systems.** Our setting is the following. We are given a finite graph $G = (V, E)$ and a finite partially ordered set S of "spin values". A *spin configuration* is an assignment of spins to vertices (sites), and our state space is the set $\mathcal{X}$ of all configurations. We are given a pmf π on $\mathcal{X}$ that is *monotone* in the sense that, when we start with π and any site v and condition on the spins at all sites other than v, the conditional law of the spin at v is monotone in the conditioning spins. We recover and (modestly) extend the Peres–Winkler result by means of the following theorem, which (i) allows somewhat more general S and (ii) encompasses—by means of Proposition 3.2, Corollary 3.3(a)–(b), and Remark 3.4—separation and L^2-distance as well as TV.

Theorem 8.3. *Fix a site v, and suppose that the conditional distributions discussed in the preceding paragraph all have positive correlations. Let K_v be the (stochastically monotone) Markov kernel for update at site v according to the conditional distributions discussed. Then we have the comparison inequality $K_v \preceq I$.*

Proof. Say that two configurations are equivalent if they differ at most in their spin at v, and let $[x]$ denote the equivalence class containing a given configuration x. Then K_v is given by

$$K_v(x, y) = \mathbf{1}(y \in [x]) \frac{\pi(y)}{\pi([x])}.$$

This K_v is the direct sum of the trivial kernels (as in Lemma 8.1) on each equivalence class. Further, each class is naturally isomorphic as a partially ordered set to S and so has positive correlations. It is well known and easily checked that K_v is stochastically monotone, so the theorem is an immediate consequence of Proposition 8.2. □

Remark 8.4. [random vs. systematic site updates] It follows [from Theorem 8.3 and Proposition 2.3(b)] for monotone spin systems with (say) linearly ordered S that, when the chains start from a common pmf having non-increasing ratio relative to π, the "systematic site updates" chain with kernel $K_{\text{syst}} := K_{v_1} \cdots K_{v_\nu}$ (for any ordering $v_1, \dots, v_\nu$ of the sites $v \in V$) mixes faster in TV, sep, and L^2 than does the "random site updates" chain with kernel $K_{\text{rand}} := \sum_{v \in V} p_v K_v$ [for any pmf $\mathbf{p} = (p_v)_{v \in V}$ on V]. This is because (recalling the paragraph preceding Proposition 2.3) the reversible kernel K_{rand} is stochastically monotone, as are K_{syst} and its time-reversal, and $K_{\text{syst}} \preceq K_{\text{rand}}$. [The explanation for the comparison here is that (as noted in the first paragraph of this section) $K_{\text{syst}} \preceq K_v$ for each $v \in V$ and (by Proposition 2.3(b)) the relation $\preceq$ on $\mathcal{K}$ is preserved under mixtures.] It is important to keep in mind here that one "sweep" of the sites using K_{syst} is counted as only one Markov-chain step.

There is a very weak ordering in the opposite direction: $K^{\nu}_{\text{rand}} \preceq pK_{\text{syst}} + (1-p)I$, with $p := \prod_{v \in V} p_v$.

8.3. **Extra updates don't delay mixing: card-shuffling.**

The following card-shuffling Markov chain, which has been studied quite a bit (see [3] and references therein) in the time-homogeneous "random updates" case where update positions are chosen independently and uniformly, is another example where comparison inequalities can be used to show that extra updates do not delay mixing.

Our state space is the set $\mathcal{X}$ of all permutations of $\{1, \dots, n\}$, and there is a parameter $p \in (0, 1)$. Given $i \in \{1, \dots, n-1\}$, we can *update* adjacent positions i and $i+1$ by sorting (i.e., putting into natural order) the two cards (numbers) in those positions with probability p and "anti-sorting" them with the remaining probability. Call the update kernel K_i. It is straightforward to check that each K_i is (i) reversible with respect to π, where $\text{inv}(x)$ is the number of inversions in the permutation x and $\pi(x)$ is proportional to $[(1-p)/p]^{\text{inv}(x)}$ [indeed, $K_i(x, \cdot)$ is the law of a permutation drawn from π but conditioned to agree with x at all positions other than i and $i+1$], and (ii) stochastically monotone with respect to the Bruhat order on $\mathcal{X}$ (defined so that $x \le y$ if y can be obtained from x by a sequence of anti-sorts of *not necessarily adjacent* cards).[7]

Theorem 8.5. *Fix a position $i \in \{1, \dots, n-1\}$, and let K_i be the Markov kernel for update of positions i and $i+1$ as discussed in the preceding paragraph. Then we have the comparison inequality $K_i \preceq I$.*

The proof of Theorem 8.5 is essentially the same as for Theorem 8.3 and therefore is omitted. The key is that the relevant equivalence classes now consist of only two permutations each and so are certainly linearly ordered, therefore having positive correlations.

8.4. **A final example.**

In a specific setting (linearly ordered state space and uniform stationary distribution) we have $K \preceq I$ quite generally:

Theorem 8.6. *Let $\mathcal{X}$ be a linearly ordered state space. If K is doubly stochastic, then $K \preceq I$ (with respect to uniform π).*

[7] To establish the monotonicity of K_i, it is sufficient to consider initial states x and y where y is obtained from x by a single anti-sort of two not necessarily adjacent cards and couple transitions from these states so that the corresponding terminal states, call them X_1 and Y_1, satisfy $X_1 \le Y_1$. A coupling that one can check works (by considering various cases) is to make the *same* decision, for x and for y, to sort or to anti-sort the cards in positions i and $i+1$.

Remark 8.7. (a) When π is uniform, to say that a kernel K is doubly stochastic is precisely to say that π is stationary for K. If K is symmetric, then Theorem 8.6 applies. Thus inserting a monotone symmetric kernel (or, more generally, a monotone doubly stochastic kernel whose transpose is also monotone) in a list of such kernels to be applied never slows mixing (by Proposition 2.4, or the more general Corollary 2.8, and the results of Section 3) when the initial pmf is non-increasing.

(b) If "linearly ordered" is relaxed to "partially ordered" in Theorem 8.6, the result is not generally true, even for monotone K. This follows from Lemma 8.1, since there are partially ordered sets for which the uniform distribution does not have positive correlations.

Proof of Theorem 8.6. We must show that $\langle Kf, g\rangle \le \langle f, g\rangle$ when f and g are nonnegative and belong to $\mathcal{M}$ (i.e., are non-increasing) and (without loss of generality) f sums to 1. It is a fundamental result in the theory of majorization [21] that f majorizes Kf if K is doubly stochastic. Since $\mathcal{X}$ is linearly ordered and f belongs to $\mathcal{M}$, it follows that, regarded as pmfs, f and Kf satisfy $Kf \ge f$ stochastically. Therefore, for $g \in \mathcal{M}$ we have $\langle Kf, g\rangle \le \langle f, g\rangle$, as desired. □

Acknowledgments. We thank Persi Diaconis, Roger Horn, Chi-Kwong Li, Guy Louchard, and Mario Ullrich for helpful discussions; an anonymous referee for helpful comments; and an anonymous associate editor for alerting us to the reference [23] (we had previously cited their work as unpublished).

References

[1] D. J. Aldous and James Allen Fill. Reversible Markov Chains and Random Walks on Graphs. Chapter drafts available from `http://www.stat.Berkeley.EDU/users/aldous/book.html`.

[2] David Aldous and Persi Diaconis. Strong uniform times and finite random walks. *Adv. in Appl. Math.*, 8(1):69–97, 1987.

[3] Itai Benjamini, Noam Berger, Christopher Hoffman, and Elchanan Mossel. Mixing times of the biased card shuffling and the asymmetric exclusion process. *Trans. Amer. Math. Soc.*, 357(8):3013–3029 (electronic), 2005.

[4] Stephen Boyd, Persi Diaconis, Pablo Parrilo, and Lin Xiao. Fastest mixing Markov chain on graphs with symmetries. *SIAM J. Optim.*, 20(2):792–819, 2009.

[5] Stephen Boyd, Persi Diaconis, Jun Sun, and Lin Xiao. Fastest mixing Markov chain on a path. *Amer. Math. Monthly*, 113(1):70–74, 2006.

[6] Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing Markov chain on a graph. *SIAM Rev.*, 46(4):667–689 (electronic), 2004.

[7] Pietro Caputo, Thomas M. Liggett, and Thomas Richthammer. Proof of Aldous' spectral gap conjecture. *J. Amer. Math. Soc.*, 23(3):831–851, 2010.

[8] Persi Diaconis and James Allen Fill. Strong stationary times via a new form of duality. *Ann. Probab.*, 18(4):1483–1522, 1990.

[9] Persi Diaconis and Laurent Saloff-Coste. Comparison theorems for reversible Markov chains. *Ann. Appl. Probab.*, 3(3):696–730, 1993.

[10] Ralf Diekmann, S. Muthukrishnan, and Madhu V. Nayakkankuppam. Engineering diffusive load balancing algorithms using experiments. In *Lecture Notes in Computer Science*, volume 1253, pages 111–122. Springer, 1997.

[11] William Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons Inc., New York, 1968.

[12] Miroslav Fiedler. Absolute algebraic connectivity of trees. *Linear and Multilinear Algebra*, 26(1-2):85–106, 1990.

[13] James Allen Fill. Bounds on the coarseness of random sums. *Ann. Probab.*, 16(4):1644–1664, 1988.

[14] James Allen Fill. Time to stationarity for a continuous-time Markov chain. *Probab. Engrg. Inform. Sci.*, 5(1):61–76, 1991.

[15] James Allen Fill. An interruptible algorithm for perfect sampling via Markov chains. *Ann. Appl. Probab.*, 8(1):131–162, 1998.

[16] James Allen Fill, Motoya Machida, Duncan J. Murdoch, and Jeffrey S. Rosenthal. Extension of Fill's perfect rejection sampling algorithm to general chains. *Random Structures Algorithms*, 17(3-4):290–316, 2000. Special issue: Proceedings of the Ninth International Conference "Random Structures and Algorithms" (Poznan, 1999).

[17] Alexander E. Holroyd. Some circumstances where extra updates can delay mixing, 2011. Preprint available: `arxiv::1101.4690v1`.

[18] Samuel Karlin and Howard M. Taylor. *A first course in stochastic processes*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, second edition, 1975.

[19] Lester H. Lange and James W. Miller. A random ladder game: permutations, eigenvalues, and convergence of Markov chains. *College Math. J.*, 23(5):373–385, 1992.

[20] László Lovász and Peter Winkler. Mixing of random walks and other diffusions on a graph. In *Survey in Combinatorics*, volume 218 of *Lecture Note Series*, pages 119 – 154. Cambridge University Press, 1995.

[21] Albert W. Marshall and Ingram Olkin. *Inequalities: theory of majorization and its applications*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1979.

[22] Yuval Peres. Lectures on "Mixing for Markov Chains and Spin Systems" (University of British Columbia, August 2005). Summary available at `http://www.stat.berkeley.edu/~peres/ubc.pdf`.

[23] Yuval Peres and Peter Winkler. Can extra updates delay mixing?, 2011. Preprint, `arXiv:1112.0603v1 [math.PR]`.

[24] James Propp and David Wilson. Coupling from the past: a user's guide. In *Microsurveys in discrete probability (Princeton, NJ, 1997)*, pages 181–192. Amer. Math. Soc., Providence, RI, 1998.

[25] James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms*, 9:223–252, 1996.

[26] James Gary Propp and David Bruce Wilson. How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *J. Algorithms*, 27(2):170–217, 1998.

[27] Sébastien Roch. Bounding fastest mixing. *Electron. Comm. Probab.*, 10:282–296 (electronic), 2005.

[28] L. Saloff-Coste and J. Zúñiga. Convergence of some time inhomogeneous Markov chains via spectral techniques. *Stochastic Process. Appl.*, 117(8):961–979, 2007.

[29] L. Saloff-Coste and J. Zúñiga. Merging for time inhomogeneous finite Markov chains. I. Singular values and stability. *Electron. J. Probab.*, 14:1456–1494, 2009.

[30] L. Saloff-Coste and J. Zúñiga. Time inhomogeneous markov chains with wave like behavior. *Ann. Appl. Probab.*, 20(5):1831–1853, 2010.

[31] L. Saloff-Coste and J. Zúñiga. Merging for inhomogeneous finite markov chains, part ii: Nash and log-sobolev inequalities. *Ann. Probab.*, to appear.

[32] Jun Sun, Stephen Boyd, Lin Xiao, and Persi Diaconis. The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Rev.*, 48(4):681–699 (electronic), 2006.

[33] David B. Wilson. Annotated bibliography of perfectly random sampling with Markov chains. In *Microsurveys in discrete probability (Princeton, NJ, 1997)*, pages 209–220. Amer. Math. Soc., Providence, RI, 1998. Latest updated version is posted at `http://dbwilson.com/exact/`.

[34] David Bruce Wilson. Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP). In *Monte Carlo methods (Toronto, ON, 1998)*, volume 26 of *Fields Inst. Commun.*, pages 143–179. Amer. Math. Soc., Providence, RI, 2000.

Department of Applied Mathematics and Statistics, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218-2682 USA
E-mail address: `jimfill@jhu.edu`

Laboratoire de Mathématiques, Université de Lille 1, Cité Scientifique – Bât. M2, 59655 Villeneuve d'Ascq CEDEX
E-mail address: `jonas.kahn@math.univ-lille1.fr`